Fast Convergence Federated Learning with Adaptive Gradient: An Application to Mental Healthcare Monitoring System

Junqiao Fan[1,3], Xuehe Wang[1(✉)], and Yuzhu Hu[2]

[1] The School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai 519082, China
wangxuehe@mail.sysu.edu.cn
[2] The School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510006, China
huyzh27@mail2.sysu.edu.cn
[3] The School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore
fanj0019@e.ntu.edu.sg

Abstract. Nowadays, there is increasing demand for mental health monitoring systems to enable disease diagnoses, such as anxiety and depression. However, the privacy concerns for sensitive data impede its wide adoption. To protect data privacy, federated learning (FL) is proposed to enable decentralized collaborative model learning without sharing sensitive data. Though, FL training process can be slowed with the non-Independent-and-Identically-Distributed (non-IID) datasets across participating clients, causing extra communication costs. In this paper, we propose the FL adaptive gradient optimization method to accelerate the convergence under the context of non-IID training. As the reference direction for parameter update, the gradient has a great impact on the convergence performance throughout the training. By adaptively modifying the local gradients according to the global gradient, we reduce the local parameter divergence to enable robust training and fast convergence. Meanwhile, as an application to our FL optimization algorithm, a novel sleep monitoring system is proposed to detect potential depression. Experiments demonstrate that with our proposed method, faster convergence and higher accuracy can be realized compared to commonly adopted Federated Averaging (FedAVG) and other adaptive optimization methods, which effectively save communication costs.

Keywords: Adaptive Gradient · Federated Learning · Non-IID Datasets · Depression Detection

1 Introduction

Since the report of the first Covid-19 case, the pandemic has spread for more than two years, overwhelming the healthcare system all over the world. The sustaining

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 255–275, 2024.
https://doi.org/10.1007/978-3-031-65126-7_24

sources and devices are employed for performance validation, and the results well demonstrate the effectiveness and advantages of the algorithm.

References

1. Chen, C.Y.: Automated ECG classification based on 1D deep learning network. Methods **202**, 127–135 (2022)
2. Ebrahimzadeh, A., Shakiba, B., Khazaee, A.: Detection of electrocardiogram signals using an efficient method. Appl. Soft Comput. **22**, 108–117 (2014)
3. Fan, X., Yao, Q., Cai, Y., Miao, F., Sun, F., Li, Y.: Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. IEEE J. Biomed. Health Inf. **22**(6), 1744–1753 (2018)
4. Guo, W., Zhang, Y., Yang, J., Yuan, X.: Re-attention for visual question answering. IEEE Trans. Image Process. **30**, 6730–6743 (2021)
5. Huang, Y., Li, H., Yu, X.: A multiview feature fusion model for heartbeat classification. Physiol. Meas. **42**(6), 065003 (2021)
6. Jekova, I., Bortolan, G., Christov, I.: Assessment and comparison of different methods for heartbeat classification. Med. Eng. Phys. **30**(2), 248–257 (2008)
7. Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J., Shao, L.: Understanding more about human and machine attention in deep neural networks. IEEE Trans. Multimedia **23**, 2086–2099 (2020)
8. Liu, P., Sun, X., Han, Y., He, Z., Zhang, W., Wu, C.: Arrhythmia classification of LSTM autoencoder based on time series anomaly detection. Biomed. Signal Process. Control **71**, 103228 (2022)
9. Mei, Z., Gu, X., Chen, H., Chen, W.: Automatic atrial fibrillation detection based on heart rate variability and spectral features. IEEE Access **6**, 53566–53575 (2018)
10. Oh, S.L., Ng, E.Y., San Tan, R., Acharya, U.R.: Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Comput. Biol. Med. **102**, 278–287 (2018)
11. Oh, S.L., Ng, E.Y., San Tan, R., Acharya, U.R.: Automated beat-wise arrhythmia diagnosis using modified U-Net on extended electrocardiographic recordings with heterogeneous arrhythmia types. Comput. Biol. Med. **105**, 92–101 (2019)
12. Pourbabaee, B., Roshtkhari, M.J., Khorasani, K.: Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Trans. Syst. Man Cybern. Syst. **48**(12), 2095–2104 (2018)
13. Mukhopadhyay, S.K., Mitra, S., Mitra, M.: An ECG signal compression technique using ASCII character encoding - sciencedirect. Measurement **45**(6), 1651–1660 (2012)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
16. Wang, J., et al.: Automated ECG classification using a non-local convolutional block attention module. Comput. Methods Programs Biomed. **203**, 106006 (2021)
17. Yeh, Y.C., Wang, W.J., Chiou, C.W.: Cardiac arrhythmia diagnosis method using linear discriminant analysis on ECG signals. Measurement **42**(5), 778–789 (2009)

Table 7. Results of ablation experiments on the 2017 PCCD

Methods	*Acc*	*Sen*	*Pre*	F1
VGG11	0.914	0.647	0.800	0.689
SEVGG11	0.972	0.901	0.898	0.898
VGG11-LSTM	0.952	0.841	0.950	0.885
VGG13-LSTM	0.945	0.868	0.893	0.880
VGG16-LSTM	0.941	0.798	0.950	0.853
SEVGG11-LSTM	0.952	0.877	0.913	0.894
Proposed (SEVGG11-LSTM-O)	**0.962**	**0.931**	**0.914**	**0.922**

4.3 Discussions

In order to determine the optimized network, convolutional networks of various depths are created. In specific, networks with 11, 13 and 16 convolutional layers are compared on the MIT-BIH arrhythmia database. As listed in Table 6, with the increasing counts of convolutional layers, the performance metrics decrease. Compared with the basic models, VGG11-LSTM is finalized for further improvement. When enhanced by the attention mechanism, the overall F1 score increases, indicating that the SEVGG11-LSTM has better classification performance. After oversampling is implemented, the classification performance is further improved, and the strong competitiveness on the MIT-BIH arrhythmia database is advantageous to that of the state-of-the-art algorithms.

Similar ablation experiments are also performed on the 2017 PCCD, and the performance records are comparatively listed in Table 7. Similarly, the F1 score decreases with the increasing counts of the convolutional layers, and is increased after introducing the attention mechanism. When oversampling is further used, the F1 score reaches the optimum value. Tables 6 and 7 well validate the great potentials of attention mechanism and oversampling for ECG classification.

Compared with the validation on a single database, two databases are employed in our paper in a "dual-centers" fashion. More participants are involved to avoid the limitation of a single database, and the conclusion is therefore more reliable. In addition, researchers can draw on the wisdom of the masses in the two databases to improve clinical trials.

5 Conclusion

This paper demonstrates a novel deep learning algorithm for ECG classification. It makes use of the combined convolutional and recurrent neural network for classifying ECG as well as the attention mechanism to assign weights. The input ECG signals are sequentially segmented and normalized. After that, the preprocessed signals are fed into the combined VGG and LSTM network for feature extraction and classification. The core network contains an attention mechanism that increases the weight of significant features. Two databases from different

Table 4. Comparison on MIT-BIH arrhythmia database

Methods	*Acc*	*Sen*	*Pre*	F1
LSTM [8]	0.986	0.980	0.976	0.978
CNN+LSTM [10]	0.981	0.975	–	–
CNN+LSTM+HOS [5]	0.989	0.965	0.969	0.967
U-Net [11]	0.973	–	–	–
RBF-BA [2]	0.952	0.956	0.906	0.930
LDA [17]	0.962	0.925	0.947	0.936
KNN [6]	–	0.809	0.769	0.788
Proposed method	**0.996**	**0.984**	**0.988**	**0.986**

Table 5. Performance comparison on 2017 PCCD

Methods	*Acc*	*Sen*	*Pre*	F1
CNN [12]	0.899	0.671	0.590	0.628
VGGNet [14]	0.976	0.909	0.903	0.906
MS-CNN [3]	0.977	0.943	0.886	0.914
BT [9]	0.966	0.832	–	–
Proposed	**0.962**	**0.931**	**0.914**	**0.922**

Table 6. Results of ablation experiments on the MIT-BIH arrhythmia database

Methods	*Acc*	*Sen*	*Pre*	F1
VGG11	0.972	0.901	0.898	0.898
VGG13	0.979	0.873	0.943	0.895
VGG16	0.963	0.810	0.858	0.824
VGG11-LSTM	0.980	0.911	0.911	0.911
VGG13-LSTM	0.980	0.894	0.919	0.906
VGG16-LSTM	0.977	0.867	0.911	0.888
SEVGG11-LSTM	0.980	0.933	0.909	0.919
Proposed (SEVGG11-LSTM-O)	**0.996**	**0.984**	**0.988**	**0.986**

Table 2. Performance on MIT-BIH arrhythmia database

Types	*Acc*	*Sen*	*Pre*	F1
A	0.994	0.955	0.955	0.955
L	0.999	0.991	1.000	0.995
N	0.992	0.997	0.989	0.993
R	1.000	1.000	1.000	1.000
V	0.994	0.978	0.997	0.987
Overall	0.996	0.984	0.988	0.986

Table 3. Performance on 2017 PCCD

Types	*Acc*	*Sen*	*Pre*	F1
N	0.962	0.974	0.982	0.978
A	0.962	0.888	0.845	0.866
Overall	0.962	0.931	0.914	0.922

4.2 Comparison Results

In order to further validate the performance of our proposal, some state-of-the-art algorithms are employed for comparison. For fair comparison, only the algorithms developed for the same classification problems and tested on the same database are introduced. Regarding the MIT-BIH arrhythmia database, the selected models for comparison are as follows. Liu *et al.* [8] proposed a model based on LSTM to obtain time-series features of ECG. In [10], the CNN layer is used to extract feature maps, and the LSTM layer captures the temporal dynamics. The model [5] is composed of a eight-layers CNN, a eight-layers LSTM and a fully connected layer. Detailed performance records of the compared models are listed in Table 4. The F1 score, *Acc*, *Sen* and *Pre* of the proposed model are higher than those of the compared models. The advantages of our proposal are therefore validated.

In addition, the two-class classification performance of wearable ECG is also compared on the 2017 PCCD. The results are listed in Table 5. Compared with the peer algorithms that have reported their F1 performance, the proposed model has the highest F1 score of 0.922. The comparison results further verify the good performance of the proposed model.

Table 1. Oversampling on MIT-BIH arrhythmia database

Type	Before oversampling	After oversampling
N	6735	6735
V	3005	6010
L	1202	7212
R	1179	7074
A	771	6939
Total	12892	33970

3.3 Cross Validation

Under the condition of limited size of dataset, 10-fold cross validation is able to achieve multiple random partitioning of the training set and test set. The original dataset is divided into 10 equal sized parts, of which nine parts are considered as training dataset and the other one is used for testing. 10-fold cross validation avoids overlap between the training and testing datasets. In each iteration, the balanced training data set is used to get the optimized parameters of the model, and then the corresponding test set is employed for performance evaluation. After 10 iterations, the results from each iteration are combined to yield the average performance of the model.

4 Results and Discussion

4.1 Overall Performance

Table 2 lists the results when using the proposed algorithm to classify various heart rhythms in the MIT-BIH arrhythmia database. The overall *Acc*, *Sen*, *Pre* and F1 values are 0.996, 0.984, 0.988, and 0.986, respectively. By considering both *Sen* and *Pre*, the F1 score better demonstrate the overall performance of the proposed algorithm. For A, L, N, R, and V types of rhythms, the achieved F1 scores of our method are 0.955, 0.995, 0.993, 1.000, and 0.987, respectively. For R rhythm, all of the performance metrics are optimal, that is, 1, which indicates the good performance in identifying R rhythms. In addition, all of the *Acc* values are higher than 0.992, demonstrating that the classification performance of our algorithm is very satisfactory.

In order to verify the universality of the proposed algorithm, it is further applied to 2017 PCCD, and the performance records are listed in Table 3. As can be observed, the *Acc*, *Sen*, *Pre* and overall F1 records are 0.962, 0.931, 0.914, and 0.922, respectively. As concluded above, the proposed algorithm shows satisfactory performance on both databases.

The fully connected layers reassembles the local features into a complete graph through the weight matrix. The representative features are integrated into a value, which has the advantage of reducing the influence of feature locations on classification results, and improving the robustness of the whole network. In addition, multiple fully connected layers are connected, so that the ability of nonlinear expression is improved.

3 Experiment Configuration

3.1 Database and Performance Metrics

Two different databases are employed for performance evaluation, the MIT-BIH arrhythmia database and the 2017 PhysioNet/CinC Challeng database (2017 PCCD). The MIT-BIH arrhythmia database includes 48 recordings, each of which is sampled at 360 Hz and contains a series of two-leads ECG data. Based on the proportion of various arrhythmias in the clinic, normal rhythm (N), left bundle branch block (R), right bundle branch block (L), ventricular precontraction (V) and atrial premature contractions (A) are selected for classification. On the other hand, there are 8528 single-lead ECG records in 2017 PCCD. All of them are collected by wearable devices at a sampling frequency of 300 Hz. The types of normal (N) and atrial fibrillation (AF) samples are selected to validate the performance of our model. Because of the data imbalance problem, samples in the training set are balanced by oversampling, as detailed in Sect. 3.2.

In this paper, accuracy (Acc), sensitivity (Sen), precision (Pre), and F1 score are employed for performance evaluation. The Acc is the proportion of correctly classified samples, Sen denotes the recognition ability of positive examples, Pre represents the proportion of correct predictions in the samples with positive predictions, and F1 score is the weighted average of Sen and Pre, respectively.

All of the models are implemented on the framework Tensorflow-gpu 2.6.2, using the Windows-11 operation system. The algorithm is deployed on a workstation with Intel(R) Core(TM) i7-12700H at 2.7 GHz, and an RTX 3060 GPU with a 14 GB memory.

3.2 Oversampling

Oversampling is adopted to solve the data imbalance problem, by increasing the number of certain arrhythmia samples. In this paper, the random oversampling method is used. It works as follows. 1) Taking the class with the largest number of samples as the benchmark, calculate the multiple of the benchmark class and the minority classes. 2) If the multiple is greater than 2, the minority classes are copied by corresponding multiples. 3) If the multiple is less than 2, a certain proportion (multiple minus one) of samples from the minority classes are randomly selected for duplication. Taking MIT-BIH arrhythmia database as an example, and samples counts before and after oversampling is listed in Table 1.

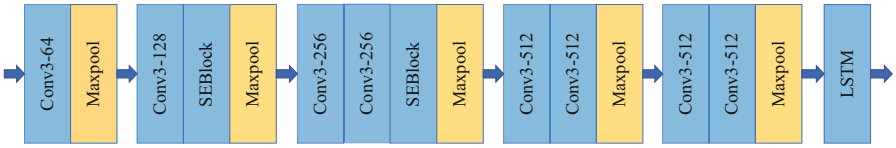


Fig. 2. Core network of the proposed model.

pooling layers. The numbers of the convolutional layers in each part are 1, 1, 2, 2, and 2, and the sizes of convolutional kernels are 64, 128, 256, 512, and 512, respectively. Each part of convolutional layer is followed by a maximum pooling layer to reduce the data length and computational burden of the model. A LSTM layer is connected after the convolutional layers and maximum pooling layers of VGGNet to avoid the gradient disappearance and gradient explosion. In addition, two SE blocks are added to the fusion model of VGGNet and LSTM, which strengthen the features of R peak. After that, the output of LSTM is given to the fully connected layer for ECG classification. During the model training process, we selected the following optimal parameters: optimizer = "Adam", learning rate = 0.001, epoch = 50, and batch size = 32 for improved performance.

The SENet, a kind of attention mechanism, is presented in the form of the SE block. It contains the squeeze and excitation modules. Squeeze operation is carried out after the traditional convolution module, that is, all spatial features in a channel are encoded into a global feature. It is defined as

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), z_c \in R^c, \tag{2}$$

where Z_c refers to the c-th element of the squeezed channels, u_c represents the c-th channel of the input, F_{sq} is the squeeze function, and H and W are the height and width of the input, respectively. After that, two fully connected layers are employed for better generalization, an activation function is used to obtain the channel-wise dependencies. The process is described by

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)), \tag{3}$$

$$\tilde{x}_c = Fscale(u_c, s_c) = s_c u_c, \tag{4}$$

where F_{ex} denotes the excitation function, W_1 and W_2 represent the widths of the inputs in dimensionality reduction and increasing layers, δ denotes the ReLU activation function, σ is the sigmoid function, and $Fscale$ represents channel-wise multiplication.

2.4 Classification Module

At the end of our model, a 3-layers fully connected layer of the VGGNet is selected, and the softmax function is used to realize multi-classification problem.

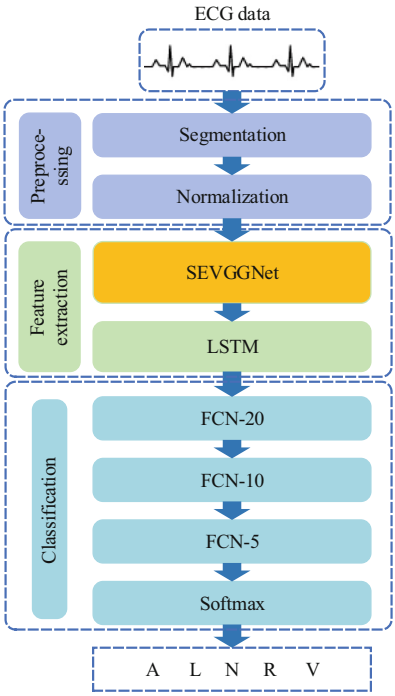


Fig. 1. Block diagram of the proposed model.

2.2 Data Preprocessing

The data preprocessing module primarily concentrates on two key tasks: signal segmentation and normalization. To alleviate computational overhead, ECG signals are segmented into multiple 10-s segments. Since ECG segment amplitudes can significantly vary owing to individual differences and lead positions, normalization becomes crucial. This normalization process involves adjusting the amplitudes, ensuring uniformity, and improving data consistency for further analysis and model training. The normalization is conducted by

$$Normalized(X) = \frac{X - \bar{X}}{S}, \tag{1}$$

where X is value of each record, $\bar{X}$ and S refer to the average and standard deviation of all the records, respectively.

2.3 Feature Extraction

The core network of the proposed algorithm is illustrated in Fig. 2. It includes eight convolutional layers, one LSTM layer, five maximum pooling layers, and two SE blocks. The convolutional layers are split into five parts by maximum

the classification performance. This advantage has been widely exploited in the fields of neural machine translation and computer vision [4,15]. It is therefore plausible to infer that adding attention mechanisms to the ECG classification model is likely to achieve performance improvement. [16] presented a CNN model with a non-local convolutional block attention module capable of distinguishing relationships between local and global segments. In addition, previous works [1,10] have demonstrated the great potentials of the combined structure of convolutional neural network (CNN) and rerrent neural network (RNN) in the ECG classification problems. As a popular CNN, the visual geometry group network (VGGNet) [14] has more nonlinear transformation and enhanced ability to learn features. On the other hand, long short-term memory neural network (LSTM) is an improved RNN, and it is able to avoid the problems of gradient explosion or gradient disappearance.

This paper proposes the SEVGGNet-LSTM model for ECG classification. Our proposal takes advantages of the combined structure of convolutional and recurrent neural network for ECG classification, and also makes full use of the attention mechanism's weight allocation capability. First, ECG data records are split into 10 s segments. After that, the amplitudes of the divided segments are normalized. The preprocessed ECG signals are then fed into the SEVGGNet-LSTM, which is a sequential combination of VGG and LSTM, with an attention mechanism (squeeze-and-excitation block, SE block) to increases the weight of important features. Finally, ECG classification is completed by the fully connected layers. Two databases from different sources and devices are employed for performance validation, and the experimental results well demonstrate the effectiveness and advantages of the proposed algorithm.

Our main contributions are as follows. 1) A fused deep convolutional neural network is constructed for ECG classification. 2) The combined convolutional and recurrent neural network and the weight allocation capability of the attention mechanism are exploited for performance promotion. 3) Two databases with ECG records collected by different devices are employed for performance evaluation.

2 The Proposed Method

2.1 Architecture

The proposed algorithm mainly consists of the preprocessing module, feature extraction module, and classification module, as illustrated in Fig. 1. The preprocessing module includes signal segmentation and normalization. The ECG signals are split into 10 s segments, after that a normalization technique is used to normalize their amplitudes. In the feature extraction module, SEVGGNet and LSTM constitute the deep neural network, and the preprocessed ECG signals are fed into the deep neural network for feature extraction. Final, a three-fully connected layer and a softmax layer constitute the classification module to realize the multi-classification problem.

SEVGGNet-LSTM: A Fused Deep Learning Model for ECG Classification

Tongyue He[1], Yiming Chen[1], Bo Fang[1], and Junxin Chen[2(✉)]

[1] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110167, China
[2] School of Software, Dalian University of Technology, Dalian 116620, China
junxinchen@ieee.org

Abstract. With the dramatic progress of smart sensing and wearable device, continuous and real-time acquisition of electrocardiograph (ECG) tends to be realized in a convenient way. Data mining of ECG signals has therefore been extensively researched, among which ECG classification is a hot topic. This paper presents a fused deep learning algorithm for ECG classification. It takes advantages of the combined convolutional and recurrent neural network for ECG classification, and the weight allocation capability of attention mechanism. The input ECG signals are firstly segmented and normalized, and then fed into the combined VGG and LSTM network for feature extraction and classification. An attention mechanism (SE block) is embedded into the core network for increasing the weight of important features. Two databases from different sources and devices are employed for performance validation, and the results well demonstrate the effectiveness and robustness of the proposed algorithm for classifying ECG signals obtained from wearable ECG devices and professional medical equipment.

Keywords: ECG classification · Deep learaning · Arrhythmia · Attention mechanism

1 Introduction

The electrocardiogram (ECG) analysis is an important non-invasive means for diagnosing and evaluating cardiac diseases [13]. However, the inherent complexity of arrhythmia often brings difficulties to medical workers in ECG classification, and may lead to mis-diagnosis. With the popularity of artificial intelligence (AI), developing computer-aided ECG classification is in a high demand.

Deep learning based solution for ECG classification has drawn world-wide concerns in recent years. It is able to automatically extract features, and hence get rid of the dependence of manual feature extraction in traditional machine learning methods. About features extraction, as reported in [7], attention mechanism is able to increase the weight of important features and further promote

Supported by the National Natural Science Foundation of China (No. 62171114).

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 245–254, 2024.
https://doi.org/10.1007/978-3-031-65126-7_23

17. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
18. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
19. Zhou, D., Wei, X.X.: Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. Adv. Neural. Inf. Process. Syst. **33**, 7234–7247 (2020)

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Alarcao, S.M., Fonseca, M.J.: Emotions recognition using EEG signals: a survey. IEEE Trans. Affect. Comput. **10**(3), 374–393 (2017)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Process. Syst. **27**, 1–9 (2014)
5. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
6. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural Inf. Process. Syst. **31**, 1–11 (2018)
7. Gao, Y., Archer, E.W., Paninski, L., Cunningham, J.P.: Linear dynamical neural population models through nonlinear embeddings. Adv. Neural Inf. Process. Syst. **29** (2016)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Jazayeri, M., Ostojic, S.: Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Curr. Opin. Neurobiol. **70**, 113–120 (2021)
10. Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. **8**, 94 (2014)
11. Pandarinath, C., et al.: Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods **15**(10), 805–815 (2018)
12. Quitadamo, L.R., et al.: Support vector machines to detect physiological patterns for EEG and EMG-based human-computer interaction: a review. J. Neural Eng. **14**(1), 011001 (2017)
13. Rossini, P.M., et al.: Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis: report from the IFCN-sponsored panel of experts. Clin. Neurophysiol. **131**(6), 1287–1310 (2020)
14. Sadtler, P.T., et al.: Neural constraints on learning. Nature **512**(7515), 423–426 (2014)
15. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 1–9 (2023)
16. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)

Table 1. Decoding accuracy from different embeddings to different labels with different decoders

Decoding Label	Emotion		Subject	
	DNN	CNN	DNN	CNN
Discovery-driven embedding-8D	0.327	0.357	0.722	0.958
Discovery-driven embedding-16D	0.355	0.355	0.746	0.962
Discovery-driven embedding-32D	0.346	0.340	0.785	**0.962**
Emotion-guided embedding-8D	0.417	0.413	0.427	0.108
Emotion-guided embedding-16D	0.420	0.410	0.357	0.091
Emotion-guided embedding-32D	0.394	0.409	0.304	0.091
Subject-guided embedding-8D	0.339	0.341	0.979	**0.995**
Subject-guided embedding-16D	0.327	0.340	0.980	**0.995**
Subject-guided embedding-32D	0.343	0.331	0.978	**0.996**

the identity of a person. Moreover, in the self-supervised learning case where no labels are provided, the accuracy of identifying a subject is still 96.2%. Recall that the input feature to generate the embedding is only EEG collected within 0.5 s, which is quite a short time. The feasibility of input features and the close-to-perfect classification accuracy indicate the promising future of our method be to applied to identify persons in various HCI scenarios.

The accuracy of detecting emotions is low, probably because the emotional changes are rarely reflected in temporal voltage features. Including more EEG features in the inputs may improve emotion detection performance.

4 Conclusion and Future Work

In this paper, we propose to use contrastive learning to generate low-dimension EEG latent embeddings that are consistent and identifiable. The contrastive learning encoder can be trained in either the supervised or self-supervised manner and the encoder trained in both manners can have very high decoding accuracy. This indicates the potential of using the EEG latent embeddings for various downstream tasks. Excitingly, we can use the EEG latent embeddings to identify different persons at close-to-perfect accuracy of 99.6% with EEG input data with only a 0.5-s window. Our method can be promisingly used in emerging modern HCI devices and applications, e.g., automatically connecting people to their roles in video games via EEG-capable AR devices.

In the future, we will try to integrate our EEG latent embedding method into industrial HCI solutions for entertainment and health management. To achieve this, we need to improve our method for a wider range of downstream applications, which may require exploring more informative EEG features as inputs. We will also verify it with various datasets and in EEG devices of different specifications.

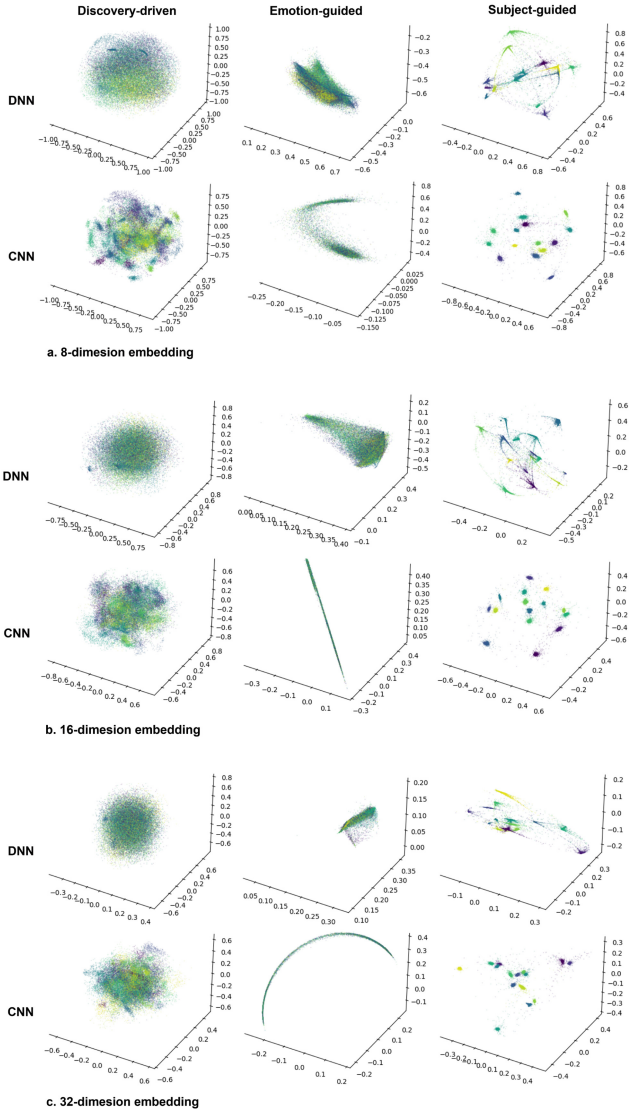


Fig. 3. The first three dimensions of DNN or CNN-learned embeddings of 8D, 16D, and 32D by discovery-driven, emotion-label-guided, and subject-label-guided contrastive learning, respectively.

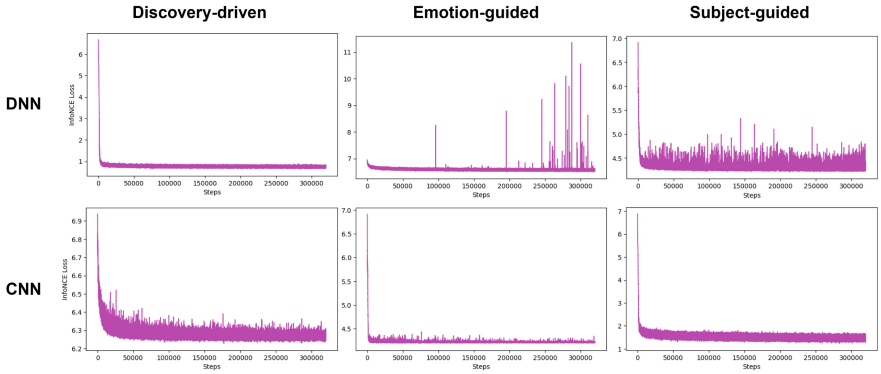


Fig. 2. The convergence performance of the DNN and CNN encoders using contrastive learning in discovery-driven, emotion-label-guided, and subject-label-guided manners, respectively

with different labels, and with different output dimensions (Fig. 3). The related labels of the data points can be distinguished by colors. From the figures, we can see that in the CNN encoder cases, where points of the same color or similar colors tend to cluster into color clouds according to some patterns, generally generate more identifiable embeddings than DNN ones, where color point clouds are more chaotic. The reason is that the CNN model considers the sequential context of the input data.

Specifically, in the CNN subject-guided cases even when the output dimension is as low as 8, we can clearly recognize 15 color clouds, which corresponds exactly to 15 subjects in the dataset. It shows that we can use a low-dimension EEG latent embedding to distinguish the identities of people, which is an exciting result and can greatly benefit new applications of HCI.

For CNN encoders trained in a discovery-driven manner, color clouds are formed in a more complex pattern and are not completely separated from each other. It indicates that the embeddings are generally informative and have great potential to be applied to various tasks. The higher the embedding dimension, the clearer the patterns and the more identifiable the color clouds.

3.3 Decoding Accuracy

The results of top-1 accuracy of decoding embeddings from different encoders to different labels are shown in Table 1. The accuracy of distinguishing different subjects has reached as high as 99.6% using 32-dimension latent embedding generated by the CNN encoder via supervised contrastive learning. Even for these latent embeddings of dimensions 8 and 16, the accuracy is as high as 99.5%, which is almost the same as the 32-dimension ones. It shows that we can compress EEG data to float vectors as small as 8 dimensions to represent

Positive and negative samples are taken from a minibatch of the training input. The identification of positive and negative samples depends on the scientific problem we are solving. In the discovery-driven manner when no label is provided, samples near x along the timeline are positive and those far away from x along the timeline are negative. While in the hypothesis manner, specific labels are provided and the learning process is similar to supervised contrastive learning in concept. Samples with the same label as that of x are positive, while those with different labels from x are negative. We train different encoder models without any label, with emotion labels, and with subject labels, respectively, and compare their convergence, latent visualization, and decoding performance, respectively.

We test the encoder with two different neural network structures, where one is a five-layer 1D convolutional network with skipping connections (CNN) and the other is a four-layer fully connected network (DNN). Perceptrons in both networks are activated by GELU functions. The mini-batch size of the input is 1,024, the learning rate is 0.001. The encoder output dimension can be 8, 16, and 32, and the number of mini-batch training iterations is 10,000 times the encoder output dimension.

Embedding Decoding. The decoder is a classification or regression model that fits the EEG latent embedding to the task-specific labels. We use the K-nearest neighbors (KNN) method (where $k = 5$) as the model and emotions and subjects as the labels, respectively. Both types of labels are discrete, where emotion labels have 3 values and subjects have 15. The embedding decoder is trained separately from the embedding encoder, where the encoder output embeddings are generated from the training input features by the well-learned encoder and the corresponding labels are the training inputs of the decoder.

3 Results and Discussions

3.1 Contrastive Learning Convergence

We explore the convergence performance of the encoder using contrastive learning in discovery-driven and hypothesis manners. Figure 2 shows the NCE loss of the DNN and CNN encoders to encode EEG to 32-dimension latent embeddings provided without labels, with emotion labels, and with subject labels, respectively. The encoder tends to converge in all cases, which indicates that the encoder will generate consistent latent embeddings for EEG features that are considered similar. The NCE losses of encoding 8-dimension and 16 dimension latent embeddings also converge but jitter in a smaller magnitude, which is not depicted here to prevent redundancy.

3.2 Embedding Visualization

Low-dimension representations easily can be visualized to help the interpretation [9]. We visualize the first three dimensions of the testing embeddings generated by encoders of different neural network structures, contrastively learned

2 Method

Model. The whole procedure of the model is depicted in Fig. 1. We use an EEG dataset of emotion detection with neural activity with continuous long-lasting stimuli (Fig. 1.a). The encoder learns from time-domain features via a neural network with contrastive learning either in a discovery-driven manner or guided by task-specific labels and outputs EEG latent embeddings (Fig. 1.b). The decoder classifies the latent embeddings into different task-specific labels (Fig. 1.c).

Dataset. We use the SEED [18] dataset, which is a widely used open dataset designed for exploring the relationship between EEG and emotions. The dataset comprises EEG data from 15 people subjects joining a 3-session testing, with each testing session stimulated by watching 15 movie clips of a total of about 3600 s, which are continuous lasting stimuli. The movie stimuli are related to 3 emotions, i.e., positive, neutral, and negative. The EEG signals are collected by 62 electrode channels, down-sampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data grouped by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

Most work extracts EEG features from the frequency-domain representation [1,10,16], but some recent work [5,6] shows that interpretable latent embeddings can be learned from time-domain representation. We further down-sample the time-domain signals to 2 Hz so that within every 0.5 s, we can extract 5 voltage features for each of the 62 channels, namely the maximum, minimum, mean, median, and standard deviation. Finally, each encoder input is a 310-dimension vector with a 0.5-s time window, the volume of the training set is $\mathbb{R}^{270855\times310}$, and that of the testing set is $\mathbb{R}^{35280\times310}$.

Contrastive Learned Embeddings. The encoder is a non-linear convolutional neural network (CNN) or deep neural network (DNN) that applies contrastive learning optimizing the NCE loss, which follows a similar procedure as in [15].

For the input features x and y, where y is a positive or negative contrastive sample of x, let $p(x)$ be the probability density function of x, $p(y|x)$ and $q(y|x)$ be the probability density function of the positive and negative samples conditioned on x, respectively. Encoding x and y can be represented by a function f with normalized outputs, and $f(x)$ and $f(y)$ are the normalized latent embeddings, respectively. We use the dot product of $f(x)$ and $f(y)$ adjusted with a temperature parameter τ as the similarity function between these two latent embeddings, which is denoted as $\psi(x, y) = f(x)^T f(y)/\tau$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{x\sim p(x), y_+\sim p(y|x)\\ y_1,y_2,\ldots,y_n\sim q(y|x)}} [-\psi(x, y_+) + log \sum_{i=1}^{n} e^{\psi(x,y_i)}].$$

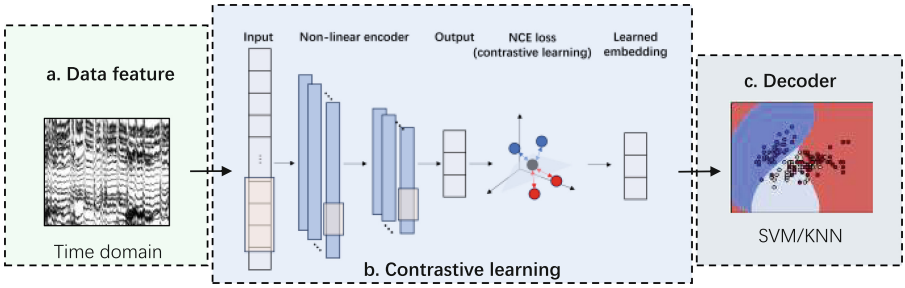


Fig. 1. The procedure of encoding time-domain EEG into embeddings by contrastive learning and decoding the learned embeddings

(HCI) and affective computing [10,16]. Encoding EEG data into consistent and identifiable latent embeddings [5,6] can greatly extend the application of EEG in various downstream tasks. The generated embeddings facilitate EEG applications by incorporating valuable information on EEG characteristics and filtering some irrelevant noise in the reduced-dimension representation. Most existing work on encoding EEG [7,14] uses supervised learning methods that depend on specific tasks, which limits its representation ability and application to other tasks. Recently, unsupervised learning [2,4,17] and self-supervised learning [3,8,11,19] methods have shown their capability of learning discriminative latent embeddings which can be generally used for downstream tasks. For example, [15] shows that contrastive learning can generate consistent and identifiable neural latent embeddings from instant spike-stimulated signals in a discovery-driven or hypothesis manner. However, neural signals are usually affected by continuous long-lasting stimuli, e.g., disease and environmental influence. The ability of contrastive learning to generate embeddings from continuous long-lasting stimulated EEG data is unknown.

In this paper, we investigate the ability of contrastive learning to generate consistent and identifiable EEG latent embeddings We use features of the time-domain representation of EEG as the input of a learnable encoder, which optimized by the noise-contrastive estimation [8] (NCE) in both the discovery-driven (self-supervised) and hypothesis (supervised) manners. We explore properties of consistency and interpretability by testing the convergence performance of the model and the visualization effect of the latent embeddings, respectively. We also verify the identifiablility of the embeddings by exploring the decoding performance of different downstream tasks. We find that encoding the time-domain features by contrastive learning can generate general EEG latent embeddings for some downstream applications. Excitingly, using the EEG latent embeddings that are encoded by 0.5-s time window features, the accuracy of recognizing the identities different persons is 96.2% in the self-supervised case and as high as 99.6% in the supervised case. The close-to-perfect decoding performance proves the potential of applying our method to emerging HCI and other EEG-related scenarios.

Contrastive Learning Consistent and Identifiable Latent Embeddings for EEG

Feng Liang[1,2(✉)], Zhen Zhang[1,2,3], Jiawei Mo[4], and Wenxin Hu[1,2]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{fliang,huwenxin}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
zhangzhen19@lzu.edu.cn
[4] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Extracting informative EEG data into low-dimension latent embeddings is important for storing and analyzing these neuron signals and applying them to various applications, such as modern human-computer interaction (HCI) techniques. We use the contrastive learning algorithm on time-domain features of EEG in both discovery-driven (self-supervised) and hypothesis (supervised) manners to encode the EEG data into latent embeddings that are proven consistent and identifiable. The self-supervised embeddings have the potential to be used for a range of downstream tasks, while the supervised embeddings have very high decoding accuracy for specific tasks. With embeddings encoded from EEG features collected within every 0.5-s window, the accuracy of recognizing the identities of persons by decoding the self-supervised and supervised embeddings is as high as 96.2% and 99.6%, respectively. Our method and results can promote new HCI techniques, e.g., automatically connecting users to their roles in AR games once they wear EEG-capable devices. The source code is available at: https://www.github.com/liangfengsid/timeEegContrastive.

Keywords: EEG · Contrastive learning · Latent embedding · Neural representation · Identity recognition

1 Introduction

Electroencephalography (EEG), the technique that intensively collects time-series electronic signals from the scalp, is widely used in clinical screening [13] and has a great potential application in modern human-computer interaction [12]

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 236–244, 2024.
https://doi.org/10.1007/978-3-031-65126-7_22

11. Li, Y., et al.: A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Trans. Cogn. Dev. Syst. **13**(2), 354–367 (2020)
12. Lin, Y.P., et al.: EEG-based emotion recognition in music listening. IEEE Trans. Biomed. Eng. **57**(7), 1798–1806 (2010)
13. Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. **8**, 94 (2014)
14. Pandarinath, C., et al.: Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods **15**(10), 805–815 (2018)
15. Pang, J.C., et al.: Geometric constraints on human brain function. Nature **618**, 566–574 (2023)
16. Rossini, P.M., et al.: Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis: report from the IFCN-sponsored panel of experts. Clin. Neurophysiol. **131**(6), 1287–1310 (2020)
17. Sadtler, P.T., et al.: Neural constraints on learning. Nature **512**(7515), 423–426 (2014)
18. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 360–368 (2023)
19. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)
20. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
21. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
22. Zhou, D., Wei, X.X.: Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. Adv. Neural. Inf. Process. Syst. **33**, 7234–7247 (2020)

5 Conclusion

In this paper, we explore encoding EEG into identifiable low-dimension latent embeddings from differential entropy powers by self-supervised contrastive learning. The latent embedding can be an informative representation uesd for downstream tasks. Using contrastive learning to extract latent embedding for EEG data is an interesting and promising topic and still needs a lot of studies. In the future, we will explore more traditional EEG features or even the raw signals for encoding EEG embeddings with contrastive learning and other self-supervised alternatives. We aim to find the algorithm to generate invariant and identifiable EEG embeddings for general tasks, and explore a wider application of EEG in the fields of neural studies, clinical screening and diagnosis, and human-computer interaction.

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
2. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural. Inf. Process. Syst. **27**, 1–9 (2014)
3. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 81–84. IEEE (2013)
4. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
5. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural. Inf. Process. Syst. **31**, 1–11 (2018)
6. Gao, Y., Archer, E.W., Paninski, L., Cunningham, J.P.: Linear dynamical neural population models through nonlinear embeddings. Adv. Neural. Inf. Process. Syst. **29** (2016)
7. Hinrikus, H., et al.: Electroencephalographic spectral asymmetry index for detection of depression. Med. Biol. Eng. Comput. **47**, 1291–1299 (2009)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Jazayeri, M., Ostojic, S.: Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Curr. Opin. Neurobiol. **70**, 113–120 (2021)
10. Khosla, P., et al.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)

Table 1. Decoding accuracy from different embeddings to different labels with different decoders

Decoding Label	Emotion		Subject
	SVM	KNN	KNN
Time nearness embedding-2D	0.338	0.382	0.040
Time nearness embedding-8D	0.359	0.398	0.044
Time nearness embedding-16D	0.415	0.428	0.065
Emotion-guided embedding-2D	0.417	0.429	0.026
Emotion-guided embedding-8D	0.500	0.447	0.054
Emotion-guided embedding-16D	**0.507**	0.464	0.072
Subject-guided embedding-2D	0.352	0.341	0.097
Subject-guided embedding-8D	0.326	0.426	**0.234**
Subject-guided embedding-16D	0.369	0.383	0.078

to extract low-dimension latent features where EEGs with similar characteristics should have similar embeddings close in distance, the purpose of the DE extraction and the contrastive learning encoder is somewhat overlapped. Besides, as EEG signals tend to vibrate in a short time and exhibit more distinguishable characteristics in a longer observation, the features extracted from short-time windows only fluctuates and may not be representative of specific properties. More input features besides DE, including statistical features about frequency domain and time domain signals and asymmetric features between electrode channels [7,11], can hopefully improve the embedding performance.

About Encoder Model. The encoder we use in this paper is a four-layer DNN. We also tested with a similar complexity CNN, alternatively. Both the encoding convergence and the visualization of the outcome embeddings are similar and the later decoding accuracy is slightly lower. We also used deeper neural networks (up to 16 1-D convolutional layers). The encoding convergence and decoding accuracy do not improve either. The reason is that the dimension of the DE feature, 310, is not large and a very deep network is not necessary. When we add more features for the encoder input, a more complex neural network may improve the embedding quality, which will be left to our future work.

About Embedding Dimension. For time-nearness-guided and emotion-guided embedding, the higher the dimension, the higher the top-1 classification accuracy for emotion labels. It indicates that high-dimension embeddings have the ability to include more useful information than low-dimension ones. But it is not the same case with subject-guided embeddings. Lower-dimension embeddings may lack representation ability, while higher-dimension embeddings may involve more noise than useful information for subject classification.

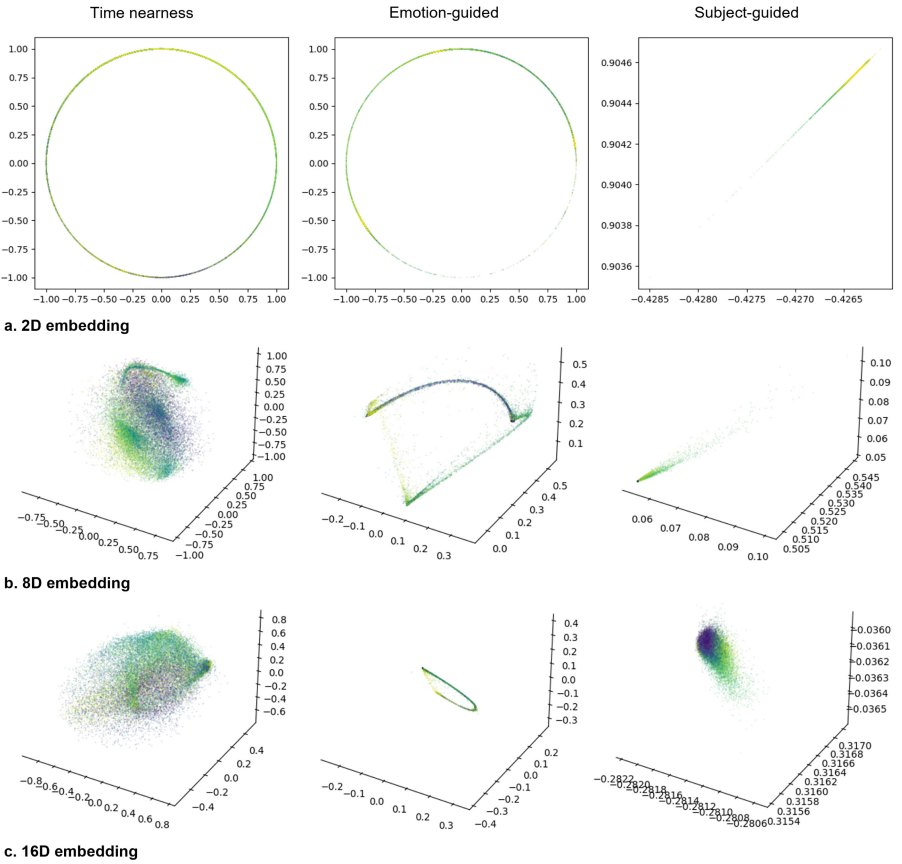


Fig. 3. The first two/three dimensions of learned embeddings of 2D, 8D, and 16D guided by time nearness, emotion labels, and subject labels, respectively.

reason is that the DE of different frequency bands varies more with people's emotional changes, but is more consistent across different people subjects. Decoding subject-guided embeddings to emotion labels also generates low accuracy, because the subject-guided embeddings have removed information about distinguishing emotions.

4 Discussion

About de Feature. The DE is proven a significant single feature which greatly reduces the feature dimension to 5 values for each channel in each short-time window, with some important frequency domain information remained and some noises filtered. But it also leaves out much useful information for encoding an invariant and identifiable embedding. Since the embedding encoder is supposed

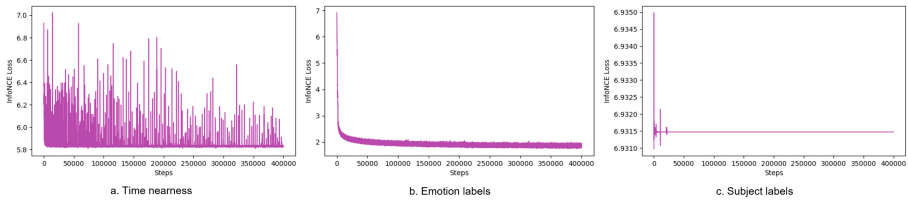


Fig. 2. NCE loss the encoder guided by different criteria

a learning rate of 0.001. The encoder does not converge in the time nearness and subject label cases. The NCE loss in the time nearness case jitters with a lower bound of about 5.8 limited by the time window length, while that in the subject label case is a straight line at the level of about 6.9315. The NCE loss in the emotion label case drops quickly in the first 1,000 iterations and gradually converges to about 1.9 after that.

3.2 Embedding Visualization

Visualizing the embeddings can help to interpret the encoding quality [9]. As shown in Fig. 3, we generate embeddings guided by different information (time or task related labels) with different output dimension sizes, 2, 8, and 16. The first two or three dimensions of the embeddings are drawn where values of the related information are indicated by colors, where embeddings related to the same information have the same color. When the embeddings are identifiable, the embedding points will cluster by color and have a clearer contour intuitively, where points of the same color are closer to each other and farther away from embedding points with different colors.

3.3 Decoding Accuracy

The results of top-1 accuracy of decoding embeddings of different dimensions from different encoder models to different labels are shown in Table 1. The highest classification accuracy for emotion labels is 0.507, which achieved by using the non-linear SVM method with 16-dimension embeddings generated by the emotion-guided encoder, and that for subject labels is 0.234, which achieved by using KNN with 8-dimension embeddings generated by the subject-guided encoder. The accuracy of decoding label-guided embeddings is higher than decoding embeddings that are guided by time nearness when no label is provided. The higher dimension of the embeddings tends to increase classification accuracy, except for decoding the subject-guided embeddings to subject labels. The embeddings generated by time nearness information can also be used for emotion classification, which indicates the potential application of contrastive learning to EEG to a wider range of downstream tasks. The classification accuracy for subject labels is much poorer than that for emotion labels. The possible

encoder is non-linear and is usually a convolutional neural network (CNN) or a deep neural network (DNN) that applies contrastive learning, which follows a similar procedure as in [18].

For the DE features h and g, where g is a positive or negative sample of h, let $p(h)$ be the probability density function of h, $p(g|h)$ and $q(g|h)$ be the probability density function of the positive and negative samples conditioned on h, respectively. After encoding h and g, $c(h)$ and $c(g)$ are their normalized latent embeddings, respectively. The similarity function between $c(h)$ and $c(g)$ is denoted as $\psi(h, g)$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{h\sim p(h), g_+\sim p(g|h)\\ g_1,g_2,\ldots,g_n\sim q(g|h)}} [-\psi(h, g_+) + log \sum_{i=1}^{n} e^{\psi(h,g_i)}].$$

Positive and negative samples are taken from a minibatch of the training input. The identification of positive and negative samples depends on the scientific problem we are solving. It can be based on time nearness between h and g if no label is provided, where samples close to h in time are considered positive and those far from h in time are considered negative. We can also provide labels to guide the training so that samples with the same label as that of h are considered positive and others are negative. The label-guided approach is supervised contrastive learning. With the SEED dataset with emotion labels from different subjects, we learn different encoder models based on time, emotion labels, and subject labels, respectively, and compare their embedding performance.

As to the similarity function, we use the dot product of the normalized latent embeddings adjusted with a temperature parameter τ, i.e., $\psi(h, g) = c(h)^T c(g)/\tau$.

Embedding Decoding. In application, EEG embeddings can be decoded for classification and regression tasks, as shown in Fig. 1.c. We use K-nearest neighbors (KNN) and non-linear support vector machine (SVM) models to classify the EEG embeddings generated by different encoders into emotion and subject labels, respectively. The embedding decoder is trained separately from the embedding encoder, and the training embeddings for the decoder are generated by the well-trained encoder from training DE features.

3 Results

3.1 Contrastive Learning Convergence

We explore the convergence performance of the encoder by contrastive learning with different criteria for identifying the positive and negative samples. Figure 2 shows the NCE loss of training a 4-layer neural network using GELU activation functions to encode EEG to 16-dimension embeddings guided by time nearness (when no label is provided), emotion labels, and subject labels, respectively. The encoder is trained for 10,000 iterations with a minibatch size of 1024 and

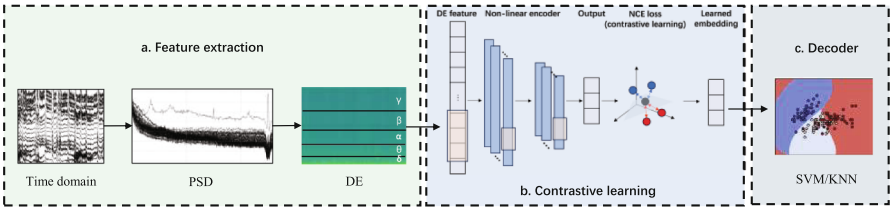


Fig. 1. The procedure of encoding EEG DE into embeddings by contrastive learning and decoding the learned embeddings

2.2 Model

The whole procedure of the model is depicted in Fig. 1. It composes three steps: the DE feature extraction, which generates DE of the frequency domain data from the origin time domain representation; the contrastive learning encoder, which encodes the DE features into latent embeddings by contrastive learning; and the decoder, which decodes the latent embedding to labels of interest.

DE Feature Extraction. DE [3] has the ability to discriminate signals between high and low frequency energy. We first transform the time domain signals to the frequency domain power in non-overlapped short-time Hanning windows and then follow a similar process to [21] to extract the DE of different frequency bands in each electrode channel (Fig. 1.a). The difference is, instead of using the magnitude spectrum as the input, we use the power spectrum density (PSD), which is recognized better than the magnitude spectrum for analyzing random vibration signals as its value is independent of frequency.

If we assume the PSD within a specific frequency band in the electrode channel i, represented by X_i, follows Gaussian distribution, i.e., $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the DE is calculated as

$$\begin{aligned} h(X_i) &= -\int_{X_i} f(x) log(f(x)) dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} log \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} dx \\ &= \frac{1}{2} log 2\pi e\sigma^2, \end{aligned}$$

where $f(x)$ is the probability density of $x \in X_i$. For each electrode channel, we divide the frequency into five bands (delta $\in [1, 4)$ Hz, theta $\in [4, 8)$ Hz, alpha $\in [8, 14)$ Hz, beta $\in [14, 31)$ Hz, and gamma $\in [31, 50)$ Hz), and calculate the DE for each frequency band, respectively. Therefore, for each short-time window, we extract 62×5 DE features, which is represented as $h(X)$.

Contrastive Learned Embeddings. As shown in Fig. 1.b, DE features are fed into a learnable encoder and the output is the EEG latent embedding. The

with specific properties of interest, much work has been done on extracting various EEG features and using different statistical or machine-learning models to retrieve useful information about behavior or health status.

Using proper features or latent embeddings of EEG is critical for EEG analysis. EEG signals usually have large sizes and come with significant noises. Most existing work uses frequency domain signals as features [13,19]. For example, it has been proved that differential entropy (DE) of different bands [3,21] incorporates useful emotional information. The latent approach [4,5] extracts invariant and identifiable latent embeddings of EEG, which can significantly reduce the representation size and extract useful information out of noises. Most work [6,17] uses supervised learning models to get latent embeddings of EEG for tasks with specific outcome labels. But much EEG data such as clinical EEG are recorded without labels, where supervised learning cannot be applied. Methods to generate general EEG latent embeddings that are independent of specific tasks can benefit the application of EEG to a wide range of downstream tasks. Recently, some studies [8,14,22] have worked on self-supervised learning methods and yielded promising results. Cebra [18] indicates that contrastive learning [1,2,10,20] has a great potential to extract invariant and identifiable EEG latent embeddings, which motivates us to explore the feasibility of extracting the general EEG latent embeddings for downstream tasks.

We apply contrastive learning, a powerful self-supervised learning algorithm, to transform DE features of EEG into lower dimensional latent embeddings for downstream tasks. We retrieve the DE of frequency power spectrum density and train a deep neural network by contrastive learning which minimizes the noise-contrastive estimation (NCE) [8] loss between generated latent embeddings of samples in a batch of training data. We use either the (self-supervised) implicit time information or (supervised) specific labels to identify positive and negative samples in the NCE loss. The first case can train the model in scenarios without labels, while the learning in the second case is guided by labels and can generate latent embeddings tuned for the specific downstream task. We explore the visual representation of the latent embeddings of different output dimensions generated by encoders guided by different information and find that the embeddings can be identifiable intuitively. We also decode the embeddings in different tasks to investigate the potential to apply the embeddings to a wide range of EEG applications.

2 Method

2.1 Dataset

We use the SEED [21] dataset, which is designed for exploring the relationship between EEG and emotions. The dataset comprises EEG data from 15 people subjects joining a 3-session testing, with each testing session stimulated by watching 15 movie clips related to 3 emotional labels. The signals are collected by 62 electrodes, downsampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data divided by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

Identifiable EEG Embeddings by Contrastive Learning from Differential Entropy Features

Zhen Zhang[1,2,3], Feng Liang[1,2(✉)], Jiawei Mo[4], and Wenxin Hu[1,2]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{fliang,huwenxin}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
zhangzhen19@lzu.edu.cn
[4] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Encoding EEG data into low-dimension latent embeddings greatly facilitates data analysis and interpretation in neuroscience studies, clinical diagnosis, and human-computer interaction. But generating informative and identifiable latent embeddings that are representative of the origin EEG is not an easy mission. Contrastive learning has the potential to utilize large amounts of unlabelled EEG data and extract informative and identifiable latent embeddings for a wide range of downstream tasks. We explore the feasibility of applying the contrastive learning method to train the EEG latent encoder from the feature of differential entropy of short-time window frequency domain signals. The encoder minimizes the noise-contrastive estimation loss by comparing the embeddings with positive and negative embedding samples, where the distinction of samples is guided by time nearness information or task-specific labels. We test encoders with different output dimensions and the outcome latent embeddings can be identifiable via visualization of a few dimensions. The decoding result also shows that the embeddings preserve information about the original EEG features and can be potentially used for a wide range of downstream tasks. The source code is available at: https://www.github.com/liangfengsid/deContrastiveLearning.

Keywords: EEG · Contrastive learning · Latent embedding

1 Introduction

Electroencephalogram (EEG) are electrical signals on the scalp collected by a set of electrodes and has been widely applied in neuroscience research [15], clinical diagnosis [16], and behavior and affection analysis [11,12]. To relate EEG

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 227–235, 2024.
https://doi.org/10.1007/978-3-031-65126-7_21

6. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
7. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural Inf. Process. Syst. **31** (2018)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Niso, G., et al.: Open and reproducible neuroimaging: from study inception to publication. In: NeuroImage 119623 (2022)
10. Schirrmeister, R.T., et al.: Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. **38**(11), 5391–5420 (2017)
11. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 360–368 (2023)
12. Song, Y., Zheng, Q., Liu, B., Gao, X.: EEG conformer: convolutional transformer for EEG decoding and visualization. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 710–719 (2022)
13. Sun, R., Sohrabpour, A., Worrell, G.A., He, B.: Deep neural networks constrained by neural mass models improve electrophysiological source imaging of spatiotemporal brain dynamics. Proc. Natl. Acad. Sci. **119**(31), e2201128119 (2022)
14. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
15. Wynn, J.K., Roach, B.J., McCleery, A., Marder, S.R., Mathalon, D.H., Green, M.F.: Evaluating visual neuroplasticity with EEG in schizophrenia outpatients. Schizophr. Res. **212**, 40–46 (2019)
16. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)

Visualizing the self-supervised learned EEG embeddings can help judge the potential value of the input feature and the effect of the encoder. In the discovery-driven case, the point cloud is in chaos when the time window size is 0.05 s, but some points of the same color start to gather and form blur contours. When the time window size is 0.5 s, we can already see some overlapping color clouds. This indicates that the corresponding input features contain some valuable information and the encoder properly transforms them into identifiable embeddings. Otherwise, if the color cloud is always in chaos, either we need to try other input features, or we need to investigate the effectiveness of the encoder algorithm.

4 Discussion and Future Work

In this paper, we use time-domain EEG features as an example to show that encoding EEG into latent embeddings and visualizing them can greatly improve the interpretability and understandability of EEG and help EEG readers easily discover some patterns. But other traditional EEG features for various tasks can also apply the latent embedding visualization so that clinical doctors and neuroscientists can make a preliminary decision from the EEG results before they go on further analysis.

In the future, we will explore self-supervised and supervised encoding methods with more traditional EEG features and see how visualizing the low-dimension embeddings can help to reveal identifiable patterns. We will also develop other visualization techniques for EEG that make EEG more understandable to clinical doctors and patients, neuroscientists, and biomedical engineers.

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
3. Daud, S.N.S.S., Sudirman, R.: Pattern of EEG voltage and oscillations under stimulation of Mozart's music and white noise for visual learning process. Biomed. Signal Process. Control **85**, 104986 (2023)
4. Donoghue, T., et al.: Parameterizing neural power spectra into periodic and aperiodic components. Nat. Neurosci. **23**(12), 1655–1665 (2020)
5. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Processing Syst. **27** (2014)

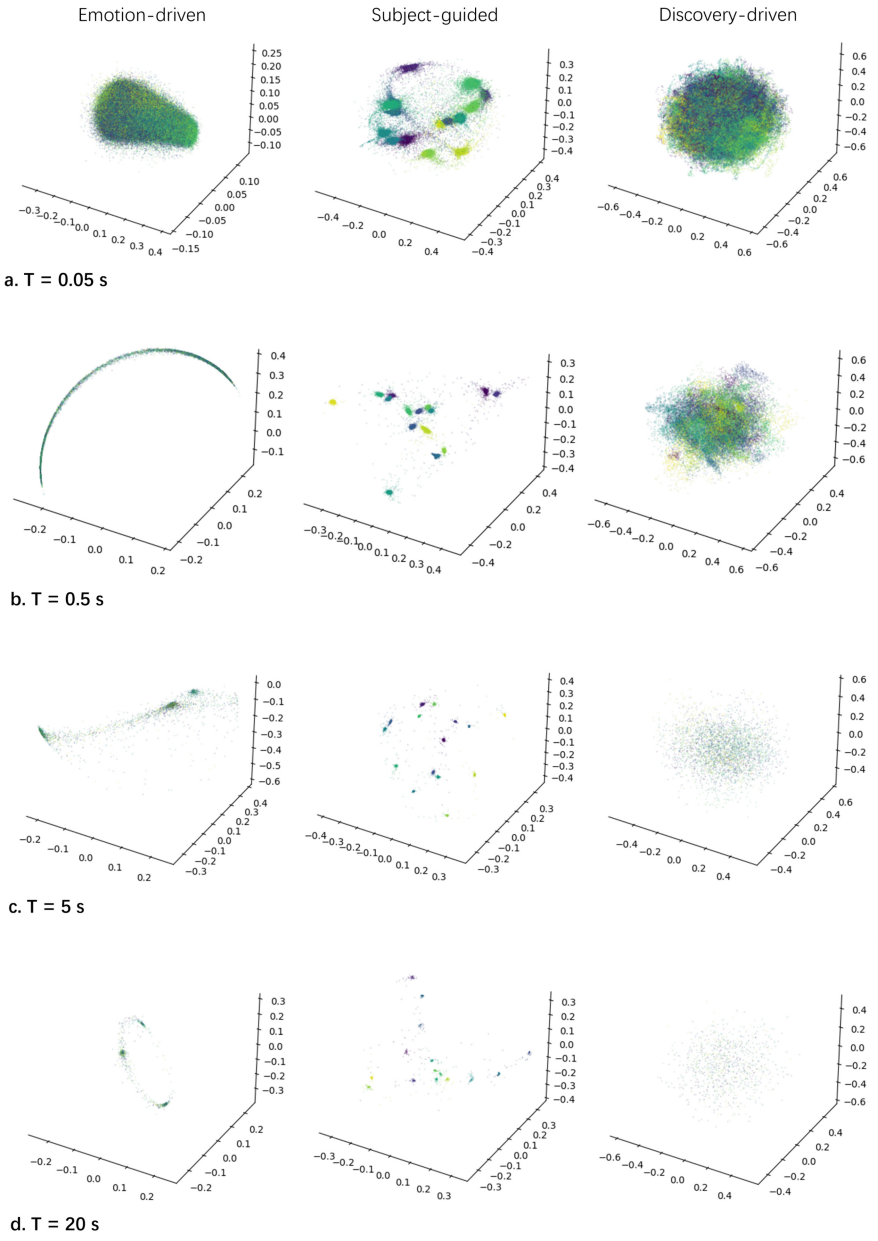


Fig. 1. The first three dimensions of EEG embeddings by discovery-driven, emotion-label-guided, and subject-label-guided contrastive learning, respectively, with features extracted from different time window sizes.

temperature parameter τ as the similarity function between these two latent embeddings, which is denoted as $\psi(x, y) = f(x)^T f(y)/\tau$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{x\sim p(x), y_+\sim p(y|x) \\ y_1, y_2, \ldots, y_n \sim q(y|x)}} [-\psi(x, y_+) + log \sum_{i=1}^{n} e^{\psi(x, y_i)}].$$

Positive and negative samples are taken from a minibatch of the training input. In the discovery-driven manner when no label is provided (self-supervised), samples near x along the timeline are positive and those far away from x along the timeline are negative. In the hypothesis manner, specific labels are provided (supervised), samples with the same label as that of x are positive, while those with different labels from x are negative. We train different encoder models without any label, with emotion labels, and with subject labels, respectively.

The encoder is a five-layer 1D convolutional network with skipping connections with each perceptron activated a GELU function. The mini-batch size of the input is 1,024, the learning rate is 0.001. The encoder output dimension is 32, and the number of mini-batch training iterations is 320,000.

2.3 Visualization

If the low-dimension EEG embeddings are invariant and discriminative, only drawing a few dimensions would be enough for revealing intuitive information about their classes. We plot the first three dimensions of the EEG embeddings of the testing set as points along the axes in the figure, with colors to indicate their related temporality or labels.

3 Results and Discussions

Figure 1 shows the results of plotting the first three dimensions of EEG embeddings by self-supervised and supervised contrastive learning methods with time-domain features extracted from different time window sizes. The clusters of color points reveal abundant information on the classes of the EEG samples.

The EEG embedding visualization can be applied to emotion recognition. In the emotion-label-guided case, we cannot find obvious patterns from the point cloud when the time window size is too small, e.g., at 0.05 or 0.5 s. But as the time window size increases, e.g., to 5 s or 20 s, points tend to cluster into 3 clouds, probably corresponding to 3 types of emotions.

Visualizing EEG embeddings can also be used for identity recognition. In the subject-label-guided case, even when the time window size is 0.05 s, the points are roughly clustered into 15 groups, probably corresponding to the 15 subjects, respectively. As the time window size increases, the contours of the point clouds become more apparent, and people can identify the cluster that a point belongs to with high confidence.

increases information density and improves interpretability when visualized [6,7,10,12]. Therefore, we are motivated to explore the visualization effect and the interpretability of EEG embeddings encoded by different methods.

We visualize the EEG embeddings generated by contrastive-learned encoders [2,8] and investigate their ability to provide distinguishable information. The contrastive-learned encoders can be either self-supervised models or supervised ones, depending on whether it is trained in the discovery-driven manner without labels provided or in the hypothesis manner with task-related labels provided. We encode different EEG features in different training manners and visualize a few dimensions of the embeddings, where the colors of points are related to the labels or their temporal information. We find that by inspecting these figures of the EEG embeddings, people can clearly identify clouds of clustered points, where each cloud consists of points whose original EEG features are considered similar. Our study shows that compressing EEG data to low-dimension embeddings by contrastive learning and visualizing only a few dimensions can help EEG readers easily recognize the inherent patterns and relationships.

2 Method

2.1 Dataset and Feature Extraction

We use the SEED [16] dataset, which comprises EEG data from 15 persons (subjects) joining a 3-session testing, with each testing session stimulated by watching 15 movie clips of a total of about 3600 s. The movie stimuli are related to 3 emotions, i.e., positive, neutral, and negative. The EEG signals are collected by 62 electrode channels, down-sampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data grouped by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

We extract time-domain features from non-overlapping sliding time windows of different sizes, i.e., 0.05, 0.5, 5, and 20 s. For every time window, we extract 5 statistic voltage features for each of the 62 channels, namely the maximum, minimum, mean, median, and standard deviation. The size of each encoder input instance is 310.

2.2 Contrastive Learned Embeddings

The encoder is a non-linear convolutional neural network (CNN) that applies contrastive learning optimizing the NCE loss, which follows a similar procedure as in [11].

For the input features x and y, where y is a positive or negative contrastive sample of x, let $p(x)$ be the probability density function of x, $p(y|x)$ and $q(y|x)$ be the probability density function of the positive and negative samples conditioned on x, respectively. Encoding x and y can be represented by a function f with normalized outputs, and $f(x)$ and $f(y)$ are the normalized latent embeddings, respectively. We use the dot product of $f(x)$ and $f(y)$ adjusted with a

Investigating the EEG Embedding by Visualization

Yongcheng Wen[1,2], Jiawei Mo[3], Wenxin Hu[1,2], and Feng Liang[1,2(✉)]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{1120200244,huwenxin,fliang}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Visualizing EEG data helps clinical doctors and neuroscientists discover potential patterns and abnormalities before further mathematical analysis. Encoding complex EEG data into low-dimension embeddings and visualizing the points in 3-dimension axes with colors can help users quickly recognize some EEG properties. We apply contrastive learning in both self-supervised and supervised manners to extract the time-domain EEG features within different time window sizes. The color points tend to cluster into clouds based on their related classes and graph readers can roughly distinguish people's emotions and identities directly by inspecting the graphs. With self-supervised encoders where the generated embeddings are supposed to be used for general tasks, the visualization method can also uncover the value of the original input features extracted from raw EEG data. The source code is available at: https://www.github.com/liangfengsid/visContrastive.

Keywords: EEG · Contrastive learning · Latent embedding · Visualization · Self-supervised

1 Introduction

Visualizing the electroencephalogram (EEG) helps users easier to discover patterns in clinical diagnosis [15] and neuroscience studies [9]. Traditional visualization elements of EEG include time-domain signals (such as the voltage) [3], frequency-domain signals (such as the power spectrum density) [4], and the source estimate [13]. However, these elements usually do not directly provide intuitively distinguishable patterns and readers may not easily extract valuable results from the visualization without further analysis. The latent approach for visualization is pervasive in computer vision [1, 5, 14]. Encoding high-dimension EEG features to low-dimension embeddings

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 221–226, 2024.
https://doi.org/10.1007/978-3-031-65126-7_20

E-Health Networks II

49. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), vol. 96, no. 34, pp. 226–231 (1996)
50. Nambiar, A., Bernardino, A., Nascimento, J.C., Fred, A.: Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In: International Conference on Automatic Face & Gesture Recognition, pp. 973–980. IEEE (2017)
51. Munaro, M., Ghidoni, S., Dizmen, D.T., Menegatti, E.: A feature-based approach to people re-identification using skeleton keypoints. In: International Conference on Robotics and Automation (ICRA), pp. 5644–5651. IEEE (2014)
52. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: International Conference on Pattern Recognition (ICPR), vol. 4, pp. 441–444. IEEE (2006)
53. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124 (2015)
54. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**(2), 207–244 (2009)
55. Davis, J.V., Kulis, B., Jain P. Sra,, S., Dhillon, I.S.: Information-theoretic metric learning. In: International Conference on Machine Learning (ICML), pp. 209–216 (2007)
56. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_21
57. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. IEEE (2010)
58. Liu, Z., Zhang, Z., Wu, Q., Wang, Y.: Enhancing person re-identification by integrating gait biometric. Neurocomputing **168**, 1144–1156 (2015)
59. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11), 2579–2605 (2008)
60. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
61. Chen, C.-H., Ramanan, D.: 3D human pose estimation= 2D pose estimation+ matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7035–7043 (2017)
62. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 30, no. 1, pp. 2058–2065 (2016)

31. Munaro, M., Basso, A., Fossati, A., Van Gool, L., Menegatti, E.: 3D reconstruction of freely moving persons for re-identification with a depth sensor. In: International Conference on Robotics and Automation (ICRA), pp. 4512–4519. IEEE (2014)
32. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1735–1742 (2006)
33. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3733–3742 (2018)
34. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
35. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Adv. Neural Inf. Process. Syst. **33**, 22243–22255 (2020)
36. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: International Conference on Learning Representation (ICLR) (2021)
37. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Process. Syst. **27**, 766–774 (2014)
38. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: International Conference on Artificial Intelligence and Statistics, pp. 297–304 (2010)
39. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 6002–6012 (2019)
40. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. arXiv preprint arXiv:1906.05849 (2019)
41. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6707–6717 (2020)
42. Ye, M., Zhang, X., Yuen, P.C., Chang, S.-F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6210–6219 (2019)
43. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 9865–9874 (2019)
44. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML), pp. 1597–1607 (2020)
45. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2020)
46. Winter, D.A.: Biomechanics and Motor Control of Human Movement. John Wiley & Sons, Hoboken (2009)
47. Aggarwal, J.K., Cai, Q., Liao, W., Sabata, B.: Nonrigid motion analysis: articulated and elastic motion. Comput. Vis. Image Underst. **70**(2), 142–156 (1998)
48. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representation (ICLR) (2018)

15. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 667–676 (2019)
16. Karianakis, N., Liu, Z., Chen, Y., Soatto, S.: Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–733. Springer, Heidelberg (2018)
17. Ge, Y., Zhu, F., Chen, D., Zhao, R.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. Adv. Neural. Inf. Process. Syst. **33**, 11309–11321 (2020)
18. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with RGB-D Sensors. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 433–442. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33863-2_43
19. Andersson, V.O., Araujo, R.M.: Person identification using anthropometric and gait data from Kinect sensor. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 425–431 (2015)
20. Rao, H., et al.: Self-supervised gait encoding with locality-aware attention for person re-identification. In: International Joint Conference on Artificial Intelligence (IJCAI), vol. 1, pp. 898–905 (2020)
21. Rao, H., et al.: A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. IEEE Trans. Pattern Anal. Mach. Intell. **01**, 1–1 (2021)
22. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 973–980 (2021)
23. Rao, H., Hu, X., Cheng, J., Hu, B.: SM-SGE: a self-supervised multi-scale skeleton graph encoding framework for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1812–1820 (2021)
24. Han, F., Reily, B., Hoff, W., Zhang, H.: Space-time representation of people based on 3D skeletal data: a review. Comput. Vis. Image Underst. **158**, 85–105 (2017)
25. Tanawongsuwan, R., Bobick, A.: Gait recognition from time-normalized joint-angle trajectories in the walking plane. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. II–II (2001)
26. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recogn. **98**, 107069 (2020)
27. Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L.: One-shot person re-identification with a consumer depth camera. In: Gong, S., Cristani, M., Yan, S., Loy, C.C. (eds.) Person Re-Identification. ACVPR, pp. 161–181. Springer, London (2014). https://doi.org/10.1007/978-1-4471-6296-4_8
28. Yoo, J.-H., Nixon, M.S., Harris, C.J.: Extracting gait signatures based on anatomical knowledge. In: Proceedings of BMVA Symposium on Advancing Biometric Technologies, pp. 596–606. Citeseer (2002)
29. Murray, M.P., Drought, A.B., Kory, R.C.: Walking patterns of normal men. J. Bone Joint Surg. **46**(2), 335–360 (1964)
30. Pala, P., Seidenari, L., Berretti, S., Del Bimbo, A.: Enhanced skeleton and face 3D data for person re-identification from depth cameras. Comput. Graph. **79**, 69–80 (2019)

Lastly, we propose a skeleton prototype contrastive learning scheme to cluster unlabeled skeleton graph representations and contrast their inherent similarity with representative skeleton features to learn effective skeleton representations for person re-ID. The proposed SPC-MGR outperforms several state-of-the-art skeleton-based methods, and is also highly effective in more general person re-ID scenarios.

References

1. Nambiar, A., Bernardino, A., Nascimento, J.C.: Gait-based person re-identification: a survey. ACM Comput. Surv. **52**(2), 33 (2019)
2. Zheng, W.-S., Gong, S., Xiang, T.: Towards open-world person re-identification by one-shot group-based verification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 591–606 (2015)
3. Baltieri, D., Vezzani, R., Cucchiara, R.: SARC3D: a new 3D body model for people tracking and re-identification. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011. LNCS, vol. 6978, pp. 107–206. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24085-0_21
4. Vezzani, R., Baltieri, D., Cucchiara, R.: People reidentification in surveillance and forensics: a survey. ACM Comput. Surv. **46**(2), 29 (2013)
5. Tan, H., Liu, X., Yin, B., Li, X.: MHSA-Net: multihead self-attention network for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8210–8224 (2022)
6. Zhu, K., Guo, H., Liu, S., Wang, J., Tang, M.: Learning semantics-consistent stripes with self-refinement for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **34**, 8531–8542 (2022)
7. Zheng, Z., Wang, X., Zheng, N., Yang, Y.: Parameter-efficient person re-identification in the 3d space. IEEE Trans. Neural Netw. Learn. Syst. **35**, 7534–7547 (2022)
8. Miao, J., Wu, Y., Yang, Y.: Identifying visible parts via pose estimation for occluded person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(9), 4624–4634 (2021)
9. Wei, Z., Yang, X., Wang, N., Gao, X.: Flexible body partition-based adversarial learning for visible infrared person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(9), 4676–4687 (2021)
10. Zhou, Q., Zhong, B., Liu, X., Ji, R.: Attention-based neural architecture search for person re-identification. IEEE Trans. Neural Netw. Learn. Syst. **33**(11), 6627–6639 (2021)
11. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE Trans. Pattern Anal. Mach. Intell. **25**(12), 1505–1518 (2003)
12. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2164–2176 (2011)
13. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. IEEE Trans. Pattern Anal. Mach. Intell. **38**(12), 2501–2514 (2016)
14. Zhao, R., Oyang, W., Wang, X.: Person re-identification by saliency learning. IEEE Trans. Pattern Anal. Mach. Intell. **39**(2), 356–370 (2017)

[58] that fuses RGB appearance and GEI features by up to 57.6% top-1 accuracy, 39.3% top-5 accuracy, and 29.1% top-10 accuracy. Despite only utilizing estimated skeleton data with noise for training, the proposed unsupervised approach can still obtain highly competitive performance compared with supervised appearance-based methods in different conditions, which demonstrates the great potential of our approach to be applied to large-scale RGB-based datasets under more general re-ID settings.

5.4 Application to Generalized Person Re-Identification

Our approach can learn a unified skeleton graph representation for different skeleton data with varying body joints or topologies, which enables the pre-trained model to be directly transferred to different datasets for the generalized person re-ID task. To evaluate the effectiveness of our approach on generalized person re-ID, we exploit the model trained on the source dataset to perform person re-ID on the target dataset, *i.e.*, direct domain generalization (DG), and then further fine-tune the model with the unlabeled data of target datasets, *i.e.*, unsupervised fine-tuning (UF), to compare the generalization performance. As shown in Table 6, we can draw the following observations and conclusions. The model trained on one dataset can be transferred to other unseen target datasets and even achieves better person re-ID performance. Direct generalization is shown to be effective among different datasets, while unsupervised fine-tuning on the target dataset can further improve the person re-ID performance. Such results demonstrate that our approach possesses good generalization ability with robustness to domain shifts [62] and can be promisingly applied to other open person re-ID tasks. Interestingly, we observe that training on different source datasets typically leads to different person re-ID performance on a new dataset. For example, the model trained on the KGBD fails to yield satisfactory performance on IAS-B, BIWI-W and BIWI-S, while the pre-trained model of KS20 with further fine-tuning on those testing sets can achieve superior performance to the original ones, as shown by the bold numbers in Table 6, which implies that an appropriate domain initialization or model pre-training of our model could be potentially exploited to facilitate better generalized person re-ID performance.

6 Conclusion

In this paper, we devise unified multi-level graphs to represent 3D skeletons, and propose an unsupervised skeleton prototype contrastive learning paradigm with multi-level relation modeling (SPC-MGR) to learn effective skeleton representations for person re-ID. We devise a multi-head structural relation layer to capture relations of neighbor body-component nodes in graphs, so as to aggregate key correlative features into effective node representations. To capture more discriminative patterns in skeletal motion, we propose a full-level collaborative relation layer to infer dynamic collaboration among different-level components. Meanwhile, a multi-level graph fusion is exploited to integrate collaborative node features across graphs to enhance structural semantics and global pattern learning.

Table 6. Generalized person re-ID performance of our approach with direct domain generalization (DG) from source datasets ("Source") to target datasets ("Target"). "UF" represents fine-tuning the source model with the unlabeled data of target datasets. BIWI-W/S denotes the Walking/Still testing set of BIWI. Bold numbers indicates that the model using "DG" or "UF" obtains better performance than the original one trained on the same dataset.

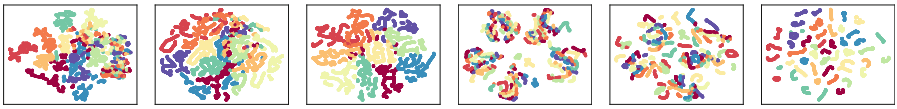
	Source	KS20				KGBD				IAS-Lab				BIWI			
Target	Type	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
KS20	**DG**	—	—	—	—	19.7	54.3	69.7	10.2	20.1	60.4	75.6	13.5	29.7	62.7	79.9	15.0
	UF	59.0	79.0	86.2	21.7	48.4	77.7	85.2	21.6	52.2	78.3	**89.1**	**22.8**	50.8	**79.7**	**87.1**	**21.9**
KGBD	**DG**	18.3	37.3	46.7	4.4	—	—	—	—	15.6	35.6	46.2	4.0	20.5	40.8	50.1	4.9
	UF	28.5	45.2	52.9	6.4	40.8	57.5	65.0	6.9	31.1	48.1	55.7	6.5	29.7	47.4	55.0	6.4
IAS-A	**DG**	27.9	57.1	71.5	15.6	29.6	56.5	71.6	16.0	—	—	—	—	27.5	53.6	67.9	14.4
	UF	**42.8**	**67.7**	**77.5**	23.0	34.1	59.4	72.0	18.4	41.9	66.3	75.6	24.2	37.2	61.8	72.2	23.8
IAS-B	**DG**	32.0	60.6	72.0	15.7	29.5	58.8	68.9	16.5	—	—	—	—	27.0	57.6	70.5	13.5
	UF	**45.4**	**68.8**	**81.4**	**29.0**	35.9	61.6	72.1	21.9	43.3	68.4	79.4	24.1	39.1	66.9	74.5	**24.2**
BIWI-W	**DG**	19.3	31.5	38.9	**19.6**	10.3	22.5	33.1	12.0	10.0	23.1	31.1	15.9	—	—	—	—
	UF	**21.6**	**32.3**	**40.9**	**20.7**	15.0	29.6	39.1	14.3	17.7	29.3	36.2	16.6	18.9	31.5	40.5	19.4
BIWI-S	**DG**	23.8	52.2	**69.9**	14.2	19.0	49.4	67.4	8.5	18.8	46.1	61.1	12.2	—	—	—	—
	UF	**40.4**	**62.9**	**74.2**	**16.2**	21.3	46.5	57.2	9.8	27.9	44.5	63.9	12.8	34.1	57.3	69.8	16.0

Such results imply that our model may learn skeleton representations with finer separation and enable pattern-based grouping in a specific class.

5.3 Application to Model-Estimated Skeleton Data

To verify the effectiveness of our skeleton-based approach when applied to large-scale RGB-based settings (CASIA-B), we exploit pre-trained pose estimation models [60,61] to extract 3D skeleton data from RGB videos of CASIA-B, and evaluate the performance of our approach with the estimated skeleton data. We compare our approach with representative appearance-based methods [54–58] and skeleton-based methods [20,23]. Note that as two fundamentally different data modalities, skeleton data are much smaller and less informative than appearance-based data (*e.g.*, RGB images), thus directly comparing skeleton-based methods with appearance-based methods is generally considered unfair. In our comparison, we provide results of classic and representative appearance-based methods as a performance reference.

As reported in Table 5, our approach is superior to recent skeleton-based methods SM-SGE and AGE with an evident performance gain of 7.2–50.4% top-1 accuracy and 1.1–5.6% mAP in four out of five evaluation conditions of CASIA-B, which substantiates that the proposed approach is capable of learning more discriminative skeleton representations than these methods in the case of using model-estimated skeleton data. Compared with representative classic appearance-based methods that utilize visual features, *e.g.*, RGB features and silhouettes, our skeleton-based approach still achieves the best performance in most conditions. For instance, our approach not only performs better than LMNN [54] and ITML [55] that use metric learning with different visual features (RGB and HSV colors and textures) [58], but also surpasses the score-based MLR model



(a) AGE (on BIWI) (b) SM-SGE (on BIWI) (c) Ours (on BIWI) (d) AGE (on KS20) (e) SM-SGE (on KS20) (f) Ours (on KS20)

Fig. 4. t-SNE visualization of the skeleton representations learned by AGE [20] ((a), (d)), SM-SGE [23] ((b), (e)), and our proposed SPC-MGR ((c), (f)) for the first 10 classes in BIWI and KS20 datasets. Note: Different colors indicate skeleton representations of different classes.

Table 5. Performance comparison with appearance-based and skeleton-based methods on CASIA-B. Note: "Cl-Nm" denotes the probe set under "Clothes" condition and gallery set under "Normal" condition. † refers to appearance-based methods and * represents requiring label information for training. "—" indicates no published result.

Probe-Gallery	Nm-Nm				Bg-Bg				Cl-Cl				Cl-Nm				Bg-Nm			
Methods	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
†LMNN* [54]	3.9	22.7	36.1	—	18.3	38.6	49.2	—	17.4	35.7	45.8	—	11.6	12.6	17.8	—	23.1	37.1	44.4	—
†ITML* [55]	7.5	22.2	34.2	—	19.5	26.0	33.7	—	20.1	34.4	43.3	—	10.3	24.5	36.1	—	21.8	30.4	36.3	—
†ELF* [56]	12.3	35.6	50.3	—	5.8	25.5	37.6	—	19.9	43.9	56.7	—	5.6	16.0	26.3	—	17.1	30.0	37.9	—
†SDALF [57]	4.9	27.0	41.6	—	10.2	33.5	47.2	—	16.7	42.0	56.7	—	11.6	19.4	27.6	—	22.9	30.1	36.1	—
†Score-based MLR* [58]	13.6	48.7	63.7	—	13.6	48.7	63.7	—	13.5	48.6	63.9	—	9.7	27.8	45.1	—	14.7	32.6	50.2	—
†Feature-based MLR* [58]	16.3	43.4	60.8	—	18.9	44.8	59.4	—	25.4	53.3	68.9	—	20.3	**42.6**	**56.9**	—	31.8	**53.6**	**64.1**	—
AGE [20]	20.8	29.3	34.2	3.5	37.1	56.2	67.0	9.8	35.5	54.3	65.3	9.6	14.6	33.0	42.7	3.0	**32.4**	51.2	60.1	3.9
SM-SGE [23]	50.2	73.5	81.9	6.6	26.6	49.0	59.4	9.3	27.2	51.4	63.2	9.7	10.6	26.3	35.9	3.0	16.6	36.8	47.5	3.5
SPC-MGR (Ours)	**71.2**	**88.0**	**92.8**	**9.1**	**44.3**	**66.4**	**76.4**	**11.4**	**48.3**	**71.6**	**81.6**	**11.8**	**22.4**	40.4	51.0	**4.3**	28.9	49.3	59.1	**4.6**

the proposed multi-level graphs (Id = 4) performs better than solely using single-level graph (Id = 3) with a remarkable margin of 2.1–16.7% top-1 accuracy and 0.4–5.7% mAP, which demonstrates that modeling body structure and relations at various levels with the proposed graph representations (MG) can encourage the model to learn more useful skeleton features for person re-ID. Adding FCRL further improves the overall performance in terms of both top-1 accuracy by 0.8–13.9% and mAP by 0.5–2.6% on different datasets. Such results verify our claim that combining structural and collaborative body relation learning can facilitate capturing richer features of body structure and skeleton patterns for the person re-ID task.

5.2 Visualization of Skeleton Representations

We conduct a t-SNE [59] visualization of skeleton representations for a qualitative analysis, and compare our approach with two state-of-the-art skeleton-based methods, *i.e.*, AGE [20], SM-SGE [23]. As presented in Fig. 4(c), the skeleton representations learned by our approach can form different class clusters with higher separation than AGE and SM-SGE on BIWI, which suggests the lower entropy of our representations. Interestingly, it is observed that the learned representations on KS20 are separated in small groups of the same class, as shown in Fig. 4(f), which enjoys significantly larger looseness than other two methods.

Table 4. Ablation study of our model with different components: Multi-level skeleton graphs (MG), multi-head structural relation layer (MSRL), full-level collaborative relation layer (FCRL), and skeleton prototype contrastive learning (SPC). "SG" denotes employing the single-level graph (part-level graph) and "+" indicates using the corresponding model component. "SG + MSRL" is evaluated under random model initialization without SPC.

Id	Configurations	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
		top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
1	Baseline	17.0	9.5	20.5	4.4	29.4	13.8	30.2	13.3	10.9	14.1	24.8	9.3
2	SG + MSRL	18.6	10.2	21.4	3.7	30.3	14.3	31.8	13.3	11.2	13.8	26.0	11.0
3	SG + MSRL + SPC	28.4	15.5	26.2	5.7	37.9	21.5	38.5	20.8	15.4	16.4	27.3	12.2
4	MG + MSRL + SPC	45.1	21.2	34.5	6.3	40.0	22.5	41.9	23.2	18.1	16.9	31.5	13.4
5	MG + MSRL + FCRL + SPC	59.0	21.7	40.8	6.9	41.9	24.2	43.3	24.1	18.9	19.4	34.1	16.0

further improvement and significantly outperforms existing supervised methods and other fine-tuned models ("SGELA + FT" and "SM-SGE + FT") on four of six testing sets. This demonstrates that our approach as a generic unsupervised contrastive paradigm can also be applied to more scenarios under label supervision. As here we directly fine-tune the model with a single MLP network (note that supervised learning is not the focus of this work), it is feasible to devise more effective supervised architectures to further boost its performance. In summary, considering that our approach does not require any manual annotation and can achieve highly competitive and more balanced performance with a significantly smaller size of network parameters, it can be a more general solution to skeleton-based person re-ID and related tasks.

5 Further Analysis

5.1 Ablation Study

In this section, we conduct ablation study to demonstrate the necessity of each component in the proposed approach. The skeleton sequences of concatenated joints are adopted as the baseline. As reported in Table 4, we can draw the following conclusions. The model utilizing single-level skeleton graph with MSRL (Id = 2, 3) shows higher performance than the baseline (Id = 1) that directly uses raw body-joint sequences by 0.3–8.5% top-1 accuracy and 0.7–7.7% mAP on all datasets, regardless of using SPC. Such results demonstrate the effectiveness of graph representations, as it can model richer body structural information and mine valuable body-component relations to obtain a more discriminative skeleton representation. Compared with the model without contrastive learning (Id = 2), employing SPC (Id = 3) obtains consistent re-ID performance improvement by up to 9.8% top-1 accuracy and 7.5% mAP on different datasets. This justifies that the proposed SPC is a highly effective contrastive learning paradigm, which enables the model to mine more typical and unique skeleton features of different identities from the unlabeled graph representations for person re-ID. Exploiting

sets of BIWI, both SGELA and the proposed approach obtain comparable mAP, while our SPC-MGR can achieve superior overall performance with higher top-1 (7.2-8.3%), top-5 (5.5-17.5%), and top-10 accuracy (5.4-25.8%). Finally, our approach also performs better than the latest graph-based skeleton representation learning method SM-SGE [23] by a distinct margin of 2.6-13.1% top-1 accuracy and 2.5-12.2% mAP on all datasets. In contrast to direct inter-sequence contrastive learning [21] or manually devising pretext tasks for skeleton representation learning [20,23], our approach can automatically mine most representative skeleton features by contrasting sequence-level representations (instances) and cluster-level representations (prototypes), which enables our model to learn better skeleton representations for person re-ID. Moreover, our model requires only 0.01M parameters and evidently lower computational complexity for skeleton representation learning compared with existing self-supervised and unsupervised methods, as shown in Table 1, which demonstrates its superior efficiency for person re-ID tasks.

We also compare the performance of our approach with state-of-the-art skeleton-based counterparts with the cross-view evaluation (CVE) setup of KS20. As shown in Table 3, our approach remarkably outperforms the latest self-supervised and multi-view skeleton-based methods SGELA [21] and SM-SGE [23] by an average margin of 6.5–32.1% top-1 accuracy, 12.8–41.0% top-5 accuracy, 17.5–40.0% top-10 accuracy, and 5.2–22.8% mAP on 24 out of 25 testing combinations of probe views and gallery views, which demonstrates that our model can learn more discriminative skeleton representations with better robustness against viewpoint variations for cross-view person re-ID.

Comparison with Hand-Crafted and Supervised Methods. Compared with D_{13} [27] and D_{16} [30] that extract hand-crafted geometric and anthropometric skeleton descriptors, our model achieves a significant improvement of person re-ID performance by 1.5–23.8% top-1 accuracy on KGBD, BIWI-S, and BIWI-W. Despite gaining similar performance on IAS-A and IAS-B, these methods are inferior to our approach by at least 7.3% top-1 accuracy on more challenging datasets such as KS20 and KGBD that contains more viewpoints and individuals. Furthermore, with *unlabeled* 3D skeletons as the only input, the proposed approach can obtain comparable or even superior performance to two state-of-the-art supervised methods PoseGait [26] and MG-SCR [22] on five out of six testing sets (KS20, IAS-A, IAS-B, BIWI-S, BIWI-W). Interestingly, with skeleton labels as the supervision, these methods still fail to obtain satisfactory person re-ID accuracy and even perform worse than hand-crafted methods on datasets with frequent view, shape, and appearance changes in KS20, IAS, and BIWI testing sets. This might also suggest that a limited amount of labeled skeleton data in small datasets such as IAS could reduce the ability of supervised models to learn discriminative features, while training with larger-scale skeleton data (KGBD) can encourage them to achieve better performance than conventional methods. On the other hand, when employing supervised fine-tuning with skeleton labels, the performance of our approach ("SPG-MGR + FT") gains a

Table 3. Performance comparison with state-of-the-art self-supervised and unsupervised methods with cross-view evaluation (CVE) setup of KS20 datasets. $0°, 30°, 90°, 130°$, and $180°$ denote probe or gallery sets in different views.

	Gallery	0°				30°				90°				150°				180°			
Probe		top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
0°	AGE [20]	46.7	74.2	83.5	22.5	11.0	35.7	47.5	10.0	8.1	29.9	47.5	9.2	7.5	26.7	43.5	8.4	7.0	23.0	37.4	8.2
	SGELA [21]	76.2	89.6	92.8	37.1	15.1	27.3	35.1	19.9	10.1	27.5	40.9	18.2	10.7	21.5	29.3	18.0	15.4	25.8	38.0	12.6
	SM-SGE [23]	58.4	84.7	92.2	27.7	17.2	50.0	63.3	10.8	7.2	21.9	39.1	10.5	4.4	19.4	34.7	9.3	10.0	23.8	33.1	9.4
	SPC-MGR (Ours)	**78.9**	**94.1**	**97.3**	**52.9**	**26.2**	**53.1**	**71.5**	**22.9**	**39.1**	**59.8**	**71.5**	**31.4**	**30.5**	**57.4**	**72.7**	**26.6**	**27.0**	**52.0**	**66.4**	**19.9**
30°	AGE	10.1	42.8	57.8	8.8	52.3	82.7	91.5	25.0	15.0	35.6	58.5	8.8	10.1	24.2	41.8	8.1	7.8	24.2	34.3	8.3
	SGELA	13.1	19.6	22.6	19.4	70.9	88.2	91.8	40.5	11.8	24.5	36.3	16.5	6.9	22.6	31.7	15.4	9.2	15.4	22.9	13.9
	SM-SGE	18.1	48.4	65.0	11.5	60.2	82.0	89.8	28.2	12.5	27.2	35.3	10.7	7.5	23.4	33.8	10.6	8.8	27.2	39.1	10.5
	SPC-MGR (Ours)	**39.1**	**60.2**	**69.1**	**26.2**	**75.4**	**95.7**	**96.5**	**56.7**	**40.2**	**62.5**	**72.3**	**32.4**	**28.9**	**55.1**	**66.0**	**24.9**	**18.4**	**48.1**	**66.4**	**16.1**
90°	AGE	7.5	27.3	43.2	8.7	9.0	28.5	44.1	9.3	57.4	81.4	90.7	19.2	13.8	41.1	57.1	9.0	7.8	30.0	46.0	8.3
	SGELA	9.6	19.8	29.7	16.4	10.8	15.6	20.4	17.5	48.4	75.7	86.5	31.6	17.1	35.7	43.0	22.0	13.5	23.4	31.8	21.3
	SM-SGE	19.1	33.1	48.1	12.4	23.1	40.6	57.4	11.5	72.2	89.1	92.8	24.9	20.9	48.4	69.4	12.8	19.4	36.9	51.6	11.3
	SPC-MGR (Ours)	**37.5**	**67.2**	**75.0**	**26.0**	**41.8**	**65.2**	**74.2**	**32.2**	**86.7**	**98.1**	**99.2**	**63.1**	**59.0**	**82.4**	**86.3**	**40.7**	**34.8**	**62.1**	**77.0**	**24.8**
150°	AGE	6.7	21.3	34.7	8.2	7.9	23.4	38.9	8.9	15.2	35.9	54.4	9.2	45.3	70.5	82.1	18.7	11.3	37.1	50.2	8.9
	SGELA	5.8	18.8	28.0	14.2	11.6	15.5	20.7	16.8	17.6	47.1	53.2	24.5	59.6	81.5	89.1	36.8	17.0	29.8	32.5	**23.0**
	SM-SGE	8.4	24.4	37.8	10.4	12.9	26.6	36.3	10.9	24.1	53.4	66.3	12.9	64.4	85.9	95.0	25.5	17.8	40.9	59.1	12.1
	SPC-MGR (Ours)	**28.5**	**59.8**	**71.9**	**23.3**	**25.4**	**49.6**	**65.2**	**22.3**	**57.4**	**77.0**	**85.2**	**40.8**	**77.3**	**96.5**	**97.7**	**58.6**	**35.9**	**62.5**	**79.3**	**23.0**
180°	AGE	7.9	17.7	32.6	8.1	5.2	22.4	33.4	8.3	10.5	25.6	34.0	8.2	11.6	33.1	52.9	8.8	47.1	72.4	82.6	22.6
	SGELA	14.0	29.1	39.2	21.3	11.9	20.6	25.9	17.3	18.6	37.8	49.7	19.4	22.7	45.9	55.2	20.7	**74.5**	**92.7**	95.1	38.3
	SM-SGE	5.6	20.0	30.6	8.5	6.6	22.7	31.6	8.6	13.8	34.1	45.6	9.4	10.3	37.5	56.6	10.4	51.9	79.7	87.8	25.6
	SPC-MGR (Ours)	**28.5**	**53.5**	**62.5**	**21.7**	**17.2**	**37.1**	**50.0**	**18.8**	**31.6**	**53.5**	**66.8**	**30.0**	**31.3**	**56.3**	**77.3**	**26.4**	65.2	89.5	**96.5**	**45.9**

that probe sequences are matched with the gallery sequences with correct identities using different-sized gallery candidate lists. We also report Mean Average Precision (mAP) [53] to evaluate the overall performance of our approach.

4.3 Performance Comparison

In this section, we compare our approach with state-of-the-art self-supervised and unsupervised skeleton-based person re-ID methods [20,21,23] on KS20, KGBD, IAS-Lab, and BIWI datasets with different probe settings in Table 1 and Table 2. As a reference for the overall performance, we include the latest supervised skeleton-based person re-ID methods [22,26] and representative handcrafted person re-ID methods [30,31]. For deep learning based methods, we also report their model sizes, *i.e.*, amount of network parameters, and computational complexity in Table 1.

Comparison with Self-supervised and Unsupervised Methods. As presented in Table 1 and Table 2, the proposed SPC-MGR enjoys distinct advantages over existing self-supervised and unsupervised methods on all datasets. Compared with AGE model [20] that learns skeleton features based on body-joint sequence representations, our approach consistently achieves higher person re-ID performance by a large margin of 7.2 to 37.9% top-1 accuracy and 6.0 to 12.8% mAP on different datasets, which demonstrates that the proposed multi-level skeleton graph representations with structural-collaborative body relation learning are more effective on modeling discriminative skeleton features for the person re-ID task. Our approach significantly outperforms the state-of-the-art skeleton contrastive learning method SGELA [21] by up to 25.2% top-1 accuracy and 11.0% mAP on KS20, KGBD, IAS-A, and IAS-B testing sets. On two testing

4 Experiments

4.1 Experimental Setup

Datasets . We evaluate our approach on four skeleton-based person re-ID benchmarks: *KGBD* [19], *BIWI* [27], *KS20* [50], *IAS-Lab* [51], and a large-scale RGB video based multi-view gait dataset: *CASIA-B* [52]. They collect skeleton data from 164, 50, 20, 11, and 124 different individuals respectively.

Implementation Details. The numbers of nodes in the part-level, body-level, and hyper-torso-level graphs are $n_1 = 10$, $n_2 = 5$, and $n_3 = 3$ respectively. The sequence length f is set to 6 for KS20, KGBD, BIWI, and IAS-Lab and $f = 40$ for CASIA-B following previous works for a fair comparison. The node feature dimension is $D_h = 8$ and the number of structural relation heads is $m = 16$ for KGBD and $m = 8$ for other datasets. We use $\lambda_C^{a,b} = 1$ $(a, b \in \{1, 2, 3\})$ to averagely fuse multi-level graph features. For DBSCAN, we empirically use maximum distance $\epsilon = 0.6$ (KGBD, BIWI), $\epsilon = 0.8$ (KS20, IAS-Lab), $\epsilon = 0.75$ (CASIA-B), and adopt minimum amount of samples $a_{min} = 4$ for KGBD and $a_{min} = 2$ for other datasets. We empirically set the temperature τ to 0.06 (KGBD), 0.075 (CASIA-B), 0.07 (BIWI), 0.08 (KS20, IAS-Lab) for skeleton prototype contrastive learning. Experiments with each evaluation setup are repeated for multiple times and the average performance is reported. More implementation details are provided in the appendices.

4.2 Evaluation Metrics

We compute Cumulative Matching Characteristics (CMC) curve and adopts top-1, top-5, and top-10 accuracy as the quantitative metrics, which indicate ratios

Table 2. Performance comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods on IAS-A, IAS-B, and KGBD testing sets. "+ FT" denotes employing supervised fine-tuning with labels. **Bold** refers to the best cases among self-supervised/unsupervised methods, and *italic numbers* indicate the best performers among supervised methods.

		IAS-A				IAS-B				KGBD			
Types	Methods	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
Hand-crafted	D_{13} [27]	40.0	58.7	67.6	24.5	43.7	68.6	76.7	23.7	17.0	34.4	44.2	1.9
	D_{16} [30]	42.7	62.9	70.7	25.2	44.5	69.1	80.2	24.5	31.2	50.9	59.8	4.0
Supervised	PoseGait [26]	28.4	55.7	*69.2*	17.5	*28.9*	*51.6*	*62.9*	*20.8*	*50.6*	*67.0*	*72.6*	*13.9*
	SGELA [21] + FT	18.0	32.1	46.2	13.5	23.6	42.9	51.9	14.8	43.7	58.7	65.0	7.1
	MG-SCR [22]	36.4	59.6	69.5	14.1	32.4	56.5	69.4	12.9	44.0	58.7	64.6	6.9
	SM-SGE [23] + FT	38.5	63.2	73.9	15.0	44.3	68.2	77.5	14.9	43.2	58.6	64.6	7.5
	SPC-MGR (Ours) + FT	*45.1*	*68.1*	*76.2*	*25.3*	*52.0*	*77.3*	*86.0*	*30.1*	42.5	59.6	67.1	9.0
Self-supervised/ Unsupervised	AGE [20]	31.1	54.8	67.4	13.4	31.1	52.3	64.2	12.8	2.9	5.6	7.5	0.9
	SGELA [21]	16.7	30.2	44.0	13.2	22.2	40.8	50.2	14.0	38.1	53.5	60.0	4.5
	SM-SGE [23]	34.0	60.5	71.6	13.6	38.9	64.1	75.8	13.3	38.2	54.2	60.7	4.4
	SPC-MGR (Ours)	**41.9**	**66.3**	**75.6**	**24.2**	**43.3**	**68.4**	**79.4**	**24.1**	**40.8**	**57.5**	**65.0**	**6.9**

k^{th} prototype $\boldsymbol{P}^k$, and τ represents the temperature for contrastive learning, where higher value of τ produces a softer probability distribution over prototypes and retains more similar information among clusters. The proposed SPC loss is essentially a generalized contrastive learning loss that combines multi-level skeleton graph modeling (see Sec. 3.1) and structural-collaborative relational feature fusion (see Sec. 3.2). We can theoretically formulate the objective of SPC as an Expectation-Maximization (EM) solution and extend it to other forms of contrastive paradigms. The theoretical analyses of SPC effectiveness and its relations to existing contrastive losses are provided in the appendices.

3.4 The Entire Approach

The computation flow of the proposed approach can be described as: $\boldsymbol{S} \rightarrow \mathcal{G}$ (Sect. 3.1) $\rightarrow \boldsymbol{F}$ (Sect. 3.2) $\rightarrow \boldsymbol{M}$ (Eq. 7) $\rightarrow \overline{\boldsymbol{M}}$ (Sect. 3.3) $\rightarrow \boldsymbol{P}$ (Eq. 10). For convenience, we use the embedding function $\psi(\cdot)$ to represent the multi-level skeleton graph representation encoding process, which can be formulated as $\psi(\boldsymbol{S}) = \overline{\boldsymbol{M}}$. We perform skeleton prototype contrastive learning by minimizing $\mathcal{L}_{\text{SPC}}$, so as to optimize $\psi(\cdot)$ and learn effective skeleton representations in an unsupervised manner. To facilitate better skeleton representation learning with more reliable clusters, we optimize our model by alternating clustering and contrastive learning. For the person re-ID task, we exploit the learned embedding function $\psi(\cdot)$ to encode each skeleton sequence of the probe set Φ_p into corresponding multi-level graph representation, $\{\overline{\boldsymbol{M}}^{p,i}\}_{i=1}^{N_2}$, and match it with representations, $\{\overline{\boldsymbol{M}}^{g,j}\}_{j=1}^{N_3}$, of the same identity in the gallery set Φ_g using Euclidean distance.

Table 1. Performance comparison with existing hand-crafted, supervised, self-supervised, and unsupervised methods on KS20, BIWI-Still (BIWI-S), and BIWI-Walking (BIWI-W) testing sets. "+ FT" denotes employing supervised fine-tuning with labels. The amount of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) for the deep learning based methods are also reported. **Bold** refers to the best cases among self-supervised/unsupervised methods, and *italic numbers* indicate the best performers among supervised methods.

				KS20				BIWI-S				BIWI-W			
Types	Methods	# Params	GFLOPs	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP	top-1	top-5	top-10	mAP
Hand-crafed	D_{13} [27]	—	—	39.4	71.7	81.7	18.9	28.3	53.1	65.9	13.1	14.2	20.6	23.7	17.2
	D_{16} [30]	—	—	51.7	77.1	86.9	24.0	32.6	55.7	68.3	16.7	17.0	25.3	29.6	18.8
Supervised	PoseGait [26]	8.93M	121.60	49.4	80.9	*90.2*	23.5	14.0	40.7	56.7	9.9	8.8	23.0	31.2	11.1
	SGELA [21] + FT	9.09M	7.48	49.7	67.0	77.1	22.2	29.2	65.2	73.8	23.5	13.9	15.3	16.7	22.9
	MG-SCR [22]	0.35M	6.60	46.3	75.4	84.0	10.4	20.1	46.9	64.1	7.6	10.8	20.3	29.4	11.9
	SM-SGE [23] + FT	6.25M	23.92	49.8	78.1	85.2	11.7	34.8	60.6	71.5	12.8	16.7	31.0	40.2	18.7
	SPC-MGR (Ours) + FT	0.03M	0.22	*65.8*	*82.4*	87.3	*30.6*	*43.8*	*73.6*	*80.5*	*20.3*	*21.5*	*33.9*	*41.0*	*22.9*
Self-supervised/Unsupervised	AGE [20]	7.15M	37.37	43.2	70.1	80.0	8.9	25.1	43.1	61.6	8.9	11.7	21.4	27.3	12.6
	SGELA [21]	8.47M	7.47	45.0	65.0	75.1	21.2	25.8	51.8	64.4	15.1	11.7	14.0	14.7	19.0
	SM-SGE [23]	5.58M	22.61	45.9	71.9	81.1	9.5	31.3	56.3	69.1	10.1	13.2	25.8	33.5	15.2
	SPC-MGR (Ours)	**0.01M**	**0.12**	**59.0**	**79.0**	**86.2**	**21.7**	**34.1**	**57.3**	**69.8**	**16.0**	**18.9**	**31.5**	**40.5**	**19.4**

Given multi-level graph representations $(\boldsymbol{M}_1, \cdots, \boldsymbol{M}_f)$ of an input skeleton sequence $(\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f))$, we first integrate graph features into a *sequence-level* skeleton graph representation:

$$\overline{\boldsymbol{M}} = \frac{1}{f} \sum_{i=1}^{f} w_i \boldsymbol{M}_i \tag{8}$$

where $\overline{\boldsymbol{M}}$ is the multi-level graph representation of skeleton sequence $\boldsymbol{S}_{1:f}$, which incorporates structural-collaborative features and temporal dynamics of f consecutive multi-level skeleton graphs, and w_i denotes the importance of i^{th} skeleton graph representation. Here we assume that each skeleton equally contributes to representing graph features of a sequence, *i.e.*, $w_i = 1$. For clarity, we use $\overline{\mathbb{M}} = \{\overline{\boldsymbol{M}}_i\}_{i=1}^{N_1}$ to represent multi-level graph representations of skeleton sequences in the training set Φ_t, which are exploited as *skeleton instances* in the proposed SPC scheme.

Then, to gather skeleton instances $\overline{\mathbb{M}}$ that contain similar features to find the representative skeleton prototypes, we leverage the DBSCAN algorithm [49], which can discover clusters with arbitrary shapes or semantics, to perform clustering as:

$$\text{DBSCAN}(\overline{\mathbb{M}}) \longrightarrow \overline{\mathbb{M}}^1, \overline{\mathbb{M}}^2, \cdots, \overline{\mathbb{M}}^z, \overline{\mathbb{M}}^o \tag{9}$$

where $\overline{\mathbb{M}} = \overline{\mathbb{M}}^1 \cup \overline{\mathbb{M}}^2 \cup \cdots \cup \overline{\mathbb{M}}^z \cup \overline{\mathbb{M}}^o$, z is the number of clusters (*i.e.*, pseudo classes), $\overline{\mathbb{M}}^k = \{\overline{\boldsymbol{M}}_i^k\}_{i=1}^{x_k}$, $k \in \{1, \cdots, z\}$, is the cluster that contains x_k instances belonging to the k^{th} pseudo class, and $\overline{\mathbb{M}}^o = \{\overline{\boldsymbol{M}}_i^o\}_{i=1}^{x_o}$ denotes the set of outlier instances that do not belong to any cluster. We compute the centroid of each cluster, which *averagely aggregates* features of skeleton instances in the cluster, to obtain corresponding skeleton prototype:

$$\boldsymbol{P}^k = \frac{1}{x_k} \sum_{i=1}^{x_k} \overline{\boldsymbol{M}}_i^k \tag{10}$$

where $\overline{\boldsymbol{M}}_i^k \in \mathbb{R}^{(n_1+n_2+n_3)\times D_h}$ is the i^{th} skeleton instance in the k^{th} cluster, and $\boldsymbol{P}^k$ denotes the k^{th} skeleton prototype.

To focus on the typical and discriminative features of skeleton prototypes as well as facilitate learning high-level skeleton semantics from different prototypes, we propose to enhance the inherent similarity of a skeleton instance to corresponding skeleton prototype and maximize its dissimilarity to other skeleton prototypes with a skeleton prototype contrastive loss as:

$$\mathcal{L}_{\text{SPC}} = \frac{1}{N} \sum_{k=1}^{z} \sum_{i=1}^{x_k} -\log \frac{\exp\left(\overline{\boldsymbol{M}}_i^k \cdot \boldsymbol{P}^k / \tau\right)}{\sum_{j=1}^{z} \exp\left(\overline{\boldsymbol{M}}_i^k \cdot \boldsymbol{P}^j / \tau\right)} \tag{11}$$

where N represents the number of all training instances, z denotes the number of skeleton prototypes, x_k is the number of skeleton instances belonging to the

Multi-level Graph Feature Fusion. To enhance structural semantics of multiple graphs (*e.g.*, global graph patterns) and adaptively integrate key correlative features in component collaboration, we exploit collaborative relations to fuse body-component node features across different spatial levels. We update the node representation ($\widehat{\boldsymbol{v}}_i^a$) of a^{th} level graph by fusing collaborative node features ($\widehat{\boldsymbol{v}}_j^b$) learned from different graphs:

$$\widehat{\boldsymbol{v}}_i^a \leftarrow \widehat{\boldsymbol{v}}_i^a + \sum_{b=a}^{3} \left(\lambda_C^{a,b} \sum_{j=1}^{n_b} \widehat{\mathbf{A}}_{i,j}^{a,b} \, \mathbf{W}_C^{a,b} \, \widehat{\boldsymbol{v}}_j^b \right) \tag{6}$$

where $\mathbf{W}_C^{a,b} \in \mathbb{R}^{D_h \times D_h}$ is a learnable weight matrix to integrate features of collaborative node $\widehat{\boldsymbol{v}}_j^b$ of b^{th} level into a^{th} level node $\widehat{\boldsymbol{v}}_i^a$. n_b denotes the number of nodes in the b^{th} level graph, and $\lambda_C^{a,b}$ represents the fusion coefficient between a^{th} level and b^{th} level graphs, which can be adjusted according to their inherent correlations (*e.g.*, level similarity). We denote the fused l^{th} level graph features of the i^{th} skeleton as $\boldsymbol{F}_i^l \in \mathbb{R}^{n_l \times D_h}$ by concatenating all node representations. Inspired by [23], we retain graph representations of each individual level and adopt their *concatenation* to represent a skeleton as follows:

$$\boldsymbol{M}_i = [\boldsymbol{F}_i^1; \boldsymbol{F}_i^2; \boldsymbol{F}_i^3] \tag{7}$$

where $\boldsymbol{M}_i \in \mathbb{R}^{(n_1+n_2+n_3) \times D_h}$ is the multi-level graph representation of the i^{th} skeleton $\boldsymbol{S}_i$, and $[;]$ indicates the concatenation of graph features. By combining all graph-level representations that integrate structural and collaborative body relation features (see Eq. 1-6), we encourage the model to capture richer features of body structure and skeleton patterns at various levels. Compared with the previous work [22] that adopts a direct graph weighting strategy, the proposed multi-level graph features fusion strategy is *learnable* and can adaptively integrate key relational features among different-level body components to enhance body and motion semantics learning.

3.3 Skeleton Prototype Contrastive Learning Scheme

As skeletons of the same individual typically share highly similar body attributes (*e.g.*, anthropometric attributes) and unique walking patterns [29], it is natural to consider mining the most *typical* attributes or patterns to identify the same person from others. To achieve this goal and encourage the model to capture more high-level skeleton semantics (*e.g.*, class-related patterns), we propose a Skeleton Prototype Contrastive learning (SPC) scheme to focus on the most *representative* skeleton graph features (referred as ***skeleton prototypes***) of pedestrians and exploit their inherent *similarity* and *dissimilarity* with other *unlabeled* graph representations (referred as ***skeleton instances***) to learn general and discriminative representations of each individual. The SPC scheme is built based on the proposed multi-level graph representations and structural-collaborative relation learning, and enables us to learn effective representations from *unlabeled* skeleton data for person re-ID.

the same computation of Eq. 3 to learn a potentially different structural relation, as shown in Fig. 3. We *averagely aggregate* features learned by m different structural relation heads as the representation of node i as follows:

$$\widehat{\boldsymbol{v}}_i^l = \frac{1}{m}\sum_{s=1}^{m}\sigma\left(\sum_{j\in\mathcal{N}_i}(\mathbf{A}_{i,j}^l)^s(\mathbf{W}_v^l)^s\boldsymbol{v}_j^l\right) \tag{4}$$

where $\widehat{\boldsymbol{v}}_i^l \in \mathbb{R}^{D_h}$ denotes the multi-head feature representation of node i in $\mathcal{G}^l$, m is the number of structural relation heads, $(\mathbf{A}_{i,j}^l)^s \in \mathbb{R}$ represents the structural relation between node i and j computed by the s^{th} structural relation head, and $(\mathbf{W}_v^l)^s$ denotes the corresponding weight matrix to perform feature mapping in the s^{th} head. Here we use *average* rather than concatenation operation to reduce feature dimension and allow for more structural relation heads. MSRL enables our model to capture the relations of correlative neighbor nodes (see Eq. 1 and 2) and integrates key spatial features into node representations of each graph (see Eq. 3 and 4). However, it only considers the local relations of the same-level components in graphs and is insufficient to capture global collaboration between different level body components, which motivates us to propose the full-level collaborative relation layer.

Full-Level Collaborative Relation Layer. Motivated by the natural property of human walking, *i.e.*, gait, which could be represented by the dynamic cooperation among body joints or between different body components [29], we expect our model to infer the degree of collaboration (referred as ***collaborative relations***) among body-component nodes in multi-level graphs, so as to capture more unique and recognizable walking patterns from the motion of skeletons. For this purpose, we propose a *full-level collaborative relation layer* (FCRL) to capture relations between a node and all motion-related nodes of the *same level* and that between a node and its spatially corresponding *higher level* body component or other potential components. As shown in Fig. 2 and Fig. 3, we compute collaborative relation matrix $\widehat{\mathbf{A}}^{a,b} \in \mathbb{R}^{n_a\times n_b}$ $(a, b \in \{1, 2, 3\}, a \leq b)$ between the a^{th} level nodes $\mathcal{V}^a$ and the b^{th} level nodes $\mathcal{V}^b$ as following:

$$\widehat{\mathbf{A}}_{i,j}^{a,b} = \mathrm{softmax}_j\left(\widehat{\boldsymbol{v}}_i^{a\top}\widehat{\boldsymbol{v}}_j^b\right) = \frac{\exp\left(\widehat{\boldsymbol{v}}_i^{a\top}\widehat{\boldsymbol{v}}_j^b\right)}{\sum_{k=1}^{n_b}\exp\left(\widehat{\boldsymbol{v}}_i^{a\top}\widehat{\boldsymbol{v}}_k^b\right)} \tag{5}$$

where $\widehat{\mathbf{A}}_{i,j}^{a,b}$ is the collaborative relation between node i in $\mathcal{G}^a$ and node j in $\mathcal{G}^b$. Here we use the inner product of multi-head node feature representations (see Eq. 4) that retain key spatial information of nodes to measure the degree of collaboration. Compared with the previous work [22] that merely considers body relations between adjacent level graphs, FCRL can capture the global collaborative relations among both *adjacent* and *non-adjacent* graphs, and meanwhile provides more comprehensive collaboration inferences between a node and all potential motion-correlated nodes in the same graph.

may act collaboratively in various global patterns during motion [47]. To exploit such internal relations to mine rich body-structure features and unique motion characteristics from skeletons, we propose the multi-level structural relation layer (MSRL) and full-level collaborative relation layer (FCRL) to model the structural and collaborative relations of body components from multi-level skeleton graphs as follows.

Multi-head Structural Relation Layer. To capture latent body structural information and learn an effective representation for each body-component node in skeleton graphs, we propose to focus on features of structurally-connected neighbor nodes, which enjoy higher correlations (referred as ***structural relations***) than distant pairs. For instance, adjacent nodes usually have closer spatial positions and similar motion tendency. Therefore, we devise a *multi-head structural relation layer* (MSRL) to learn relations of neighbor nodes and aggregate the most correlative spatial features to represent each body-component node.

We first devise a basic *structural relation head* based on the graph attention mechanism [48], which can focus on more correlative neighbor nodes by assigning larger attention weights, to capture the internal relation $e^l_{i,j}$ between adjacent nodes i and j in the same graph as:

$$e^l_{i,j} = \text{LeakyReLU}\Big(\mathbf{W}_r^{l\,\mathsf{T}}\left[\mathbf{W}^l_v \boldsymbol{v}^l_i \| \mathbf{W}^l_v \boldsymbol{v}^l_j\right]\Big) \tag{1}$$

where $\mathbf{W}^l_v \in \mathbb{R}^{D \times D_h}$ denotes the weight matrix to map the l^{th} level node features $\boldsymbol{v}^l_i \in \mathbb{R}^D$ into a higher level feature space $\mathbb{R}^{D_h}$, $\mathbf{W}^l_r \in \mathbb{R}^{2D_h}$ is a learnable weight matrix to perform relation learning in the l^{th} level graph, $\|$ indicates concatenating features of two nodes, and LeakyReLU$(\cdot)$ is a non-linear activation function. Then, to learn flexible structural relations to focus on more correlative nodes, we normalize relations using the softmax function as follows:

$$\mathbf{A}^l_{i,j} = \text{softmax}_j\left(e^l_{i,j}\right) = \frac{\exp\left(e^l_{i,j}\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(e^l_{i,k}\right)} \tag{2}$$

where $\mathcal{N}_i$ denotes directly-connected neighbor nodes (including i) of node i in graph. We use structural relations $\mathbf{A}^l_{i,j}$ to aggregate features of most relevant nodes to represent node i:

$$\overline{\boldsymbol{v}}^l_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \mathbf{A}^l_{i,j} \mathbf{W}^l_v \boldsymbol{v}^l_j\right) \tag{3}$$

where $\sigma(\cdot)$ is a non-linear function and $\overline{\boldsymbol{v}}^l_i \in \mathbb{R}^{D_h}$ is feature representation of node i computed by a structural relation head.

To sufficiently capture potential structural relations (*e.g.*, position similarity and movement correlations) between each node and its neighbor nodes, we employ *multiple structural relation heads*, each of which independently executes

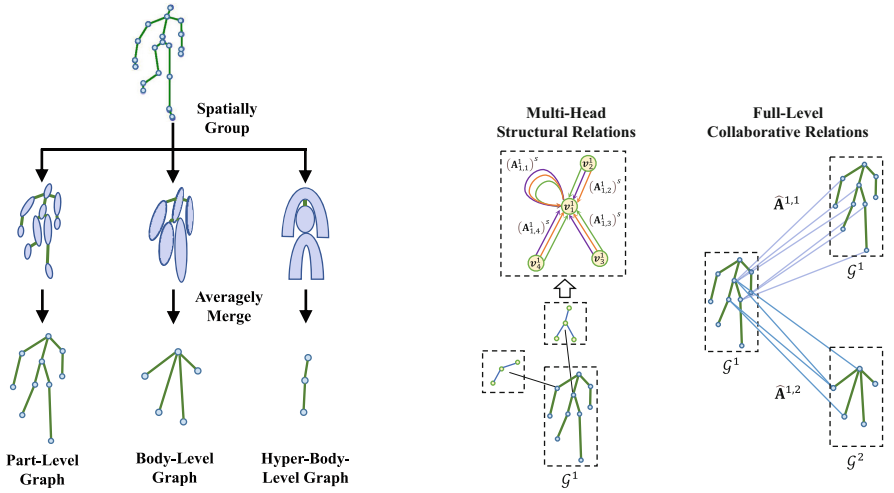


Fig. 3. Left: Three graph levels for a skeleton. We spatially divide human body into 10, 5 and 3 partitions to construct part-level, body-level, and hyper-body-level graphs, and averagely merge internal body joints into nodes. Right: Examples of multi-head structural relations in $\mathcal{G}^1$ and full-level collaborative relations among graphs ($\mathcal{G}^1$, $\mathcal{G}^1$ and $\mathcal{G}^1$, $\mathcal{G}^2$).

define a graph's adjacency matrix as $\mathbf{A}^l \in \mathbb{R}^{n_l \times n_l}$ to represent structural relations among n_l nodes. We compute the *normalized* structural relations between node i and its neighbors, *i.e.*, $\sum_{j \in \mathcal{N}_i} \mathbf{A}^l_{i,j} = 1$, where $\mathcal{N}_i$ denotes the neighbor nodes of node i. $\mathbf{A}^l$ is adaptively learned to capture flexible structural relations in the training stage.

Remarks: Compared with [22] that relies on a specific topology of original skeletons (*e.g.*, joint-level graphs), the proposed unified multi-level skeleton graphs can be viewed as *topology-independent* as they *unify* different skeleton data into an identical number of pre-defined body partitions. It can be generalized to different skeleton datasets and enables the pre-trained model to be directly transferred across different domains for generalized person re-ID (see Sect. 5.4). Besides, they can be extended to skeleton data estimated from RGB videos to learn effective person re-ID representations (see Sect. 5.3). It should be noted that the multi-level graphs can be further extended with different pre-defined body partitions. In our work, we adopt hyper-body-level, body-level, and part-level graphs since their coarse-to-fine body-component divisions match human cognition and prior knowledge of body construction [23,46].

3.2 Structural-Collaborative Body Relation Modeling

The physical connections of body structure typically endow body components in a local partition with higher correlations, while components of different parts

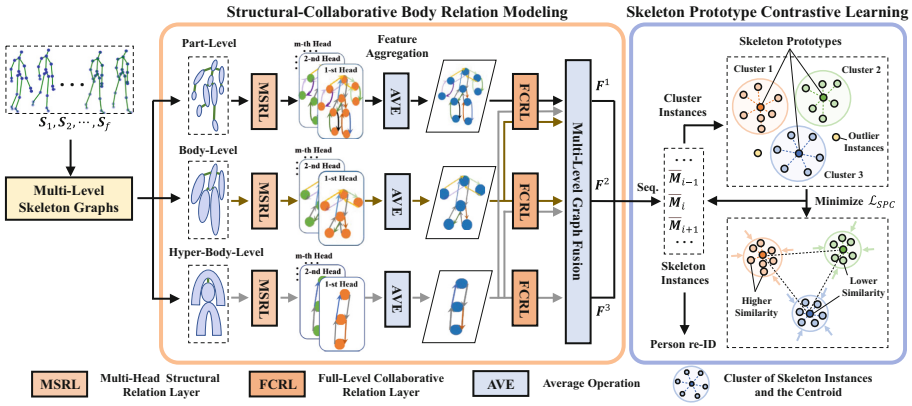


Fig. 2. Schematic diagram of SPC-MGR. Firstly, each 3D skeleton of an input sequence $\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_f$ is represented with part-level, body-level, and hyper-body-level graphs. Secondly, we employ multi-head structural relation layers (MSRL) to capture structural relations of neighbor nodes in each graph, and averagely aggregate features learned by multiple heads to obtain node representations. Then, full-level collaborative relation layers (FCRL) infer the dynamic collaborative relations among the same-level and different-level body components, which are exploited to integrate key graph features into multi-level skeleton graph representations $\boldsymbol{F}^1$, $\boldsymbol{F}^2$, and $\boldsymbol{F}^3$. Next, we perform clustering on skeleton instances, which are sequence-level ("Seq.") multi-level skeleton graph representations, to generate clusters and corresponding skeleton prototypes. Finally, during skeleton prototype contrastive learning, we enhance the similarity of instances belonging to the same prototype and maximize their dissimilarity to other prototypes by minimizing contrastive loss $\mathcal{L}_{\text{SPC}}$. The learned skeleton graph representations are exploited to perform person re-ID.

3.1 Multi-level Skeleton Graphs

Inspired by the fact that human motion can be decomposed into movements of functional body-components (*e.g.*, legs, arms) [23,46], we spatially group skeleton joints to be *higher level* body components at their centroids. Specifically, we first divide human skeletons into several partitions *from coarse to fine*. Based on the nature of body structure, we specify the location of each body partition and its corresponding skeleton joints of different sources (*e.g.*, datasets). Then, we adopt the weighted average of body joints in the same partition as the node of higher level body component and use its physical connections as edges, so as to build unified skeleton graphs for an input skeleton. As shown in Fig. 3, we construct three levels of skeleton graphs, namely *part-level*, *body-level* and *hyper-body-level* graphs for each skeleton $\boldsymbol{S}$, which can be represented as $\mathcal{G}^1$, $\mathcal{G}^2$ and $\mathcal{G}^3$ respectively. Each graph $\mathcal{G}^l(\mathcal{V}^l, \mathcal{E}^l)$ $(l \in \{1, 2, 3\})$ consists of nodes $\mathcal{V}^l = \{\boldsymbol{v}_1^l, \boldsymbol{v}_2^l, \cdots, \boldsymbol{v}_{n_l}^l\}$, $\boldsymbol{v}_i^l \in \mathbb{R}^D$, $i \in \{1, \cdots, n_l\}$ and edges $\mathcal{E}^l = \{e_{i,j}^l \mid \boldsymbol{v}_i^l, \boldsymbol{v}_j^l \in \mathcal{V}^l\}$, $e_{i,j}^l \in \mathbb{R}$. Here $\mathcal{V}^l$ and $\mathcal{E}^l$ denote the set of nodes corresponding to different body components and the set of their internal connection relations, respectively. n_l denotes the number of nodes in $\mathcal{G}^l$. More formally, we

datasets with varying skeletal topologies. Furthermore, a new full-level collaborative relation layer is devised to capture not only the cross-level relations in [22] but also more comprehensive relations among body components at the same level and non-adjacent levels, while a new learnable multi-level graph feature fusion strategy is explored to enhance graph semantics and global pattern learning.

2.2 Contrastive Learning

Contrastive learning has recently achieved great success in many self-supervised and unsupervised learning tasks [17,21,32–36]. Its general objective is to learn effective data representations by pulling closer positive pairs and pushing apart negative pairs in the feature space using contrastive losses, which are often designed based on certain auxiliary tasks (*e.g.*, similarity metrics learning). For example, Wu *et al.* [33] devise an instance-level discrimination method in the form of exemplar task [37] to perform image contrastive learning with noise-contrastive estimation loss (NCE) [38]. In [34], contrastive predictive coding (CPC) based on a probabilistic contrastive loss (InfoNCE) is proposed to learn general representations for different domains. To optimize representation learning (*e.g.*, consistency) in memory bank based contrastive methods [39–41], some recent end-to-end works [42–44] utilize all samples of the current mini-batch to generate negative instance features, while the momentum-based approach [45] further explores the use of momentum-updated encoder and queue dictionary to improve consistency of both encoder and instance features. The PCL [36] integrates both contrastive learning and clustering into an expectation-maximization (EM) framework, which is highly efficient on unsupervised visual representation learning and inspires our work for 3D skeletons.

3 The Proposed Approach

Suppose that a 3D skeleton sequence $\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f) \in \mathbb{R}^{f \times J \times D}$, where $\boldsymbol{S}_t \in \mathbb{R}^{J \times D}$ is the t^{th} skeleton with J body joints and $D = 3$ dimensions. Each skeleton sequence $\boldsymbol{S}_{1:f}$ corresponds to an ID label y, where $y \in \{1, \cdots, C\}$ and C is the number of different persons. The training set $\Phi_t = \left\{ \boldsymbol{S}_{1:f}^{t,i} \right\}_{i=1}^{N_1}$, probe set $\Phi_p = \left\{ \boldsymbol{S}_{1:f}^{p,i} \right\}_{i=1}^{N_2}$, and gallery set $\Phi_g = \left\{ \boldsymbol{S}_{1:f}^{g,i} \right\}_{i=1}^{N_3}$ contain N_1, N_2, and N_3 skeleton sequences of different persons under varying views or scenes. Our goal is to learn an embedding function $\psi(\cdot)$ that maps Φ_p and Φ_g to effective skeleton representations $\{\overline{\boldsymbol{M}}^{p,i}\}_{i=1}^{N_2}$ and $\{\overline{\boldsymbol{M}}^{g,j}\}_{j=1}^{N_3}$ *without using any label*, such that the representation $\overline{\boldsymbol{M}}^{p,i}$ in the probe set can match the representation $\overline{\boldsymbol{M}}^{g,j}$ of the same identity in the gallery set. The overview of the proposed approach is given in Fig. 2, and we present the details of each technical component below.

– We present a skeleton prototype contrastive learning (SPC) scheme based on the proposed multi-level skeleton graph representations to capture representative discriminative skeleton features and high-level class-related semantics from *unlabeled* skeleton data for person re-ID.
– Extensive experiments show that the proposed SPC-MGR outperforms several state-of-the-art skeleton-based methods on four person re-ID benchmarks, and is also highly effective when applied to skeleton data estimated from large-scale RGB videos under more general re-ID settings.

2 Related Works

2.1 Skeleton-Based Person Re-identification

Hand-Crafted Methods. Early skeleton-based works extract hand-crafted descriptors in terms of certain geometric, morphological or anthropometric attributes of human body. Barbosa *et al.* [18] compute 7 Euclidean distances between the floor plane and joint or joint pairs to construct a distance matrix, which is learned by a quasi-exhaustive strategy to extract discriminative features for person re-ID. Munaro *et al.* [27] and Pala *et al.* [30] further extend them to 13 (D_{13}) and 16 skeleton descriptors (D_{16}) respectively, and leverage support vector machine (SVM), k-nearest neighbor (KNN) or Adaboost classifiers for person re-ID. Since such solutions using 3D skeletons alone are hard to achieve satisfactory performance, they usually combine other modalities such as 3D point clouds [31] and 3D face descriptors [30] to improve person re-ID accuracy.

Supervised and Self-supervised Methods. Most recently, a few works exploit deep learning paradigms to learn gait representations from skeleton data for person re-ID in a supervised or self-supervised manner. Liao *et al.* [26] propose PoseGait, which feeds 81 hand-crafted pose features of 3D skeletons into CNN for human recognition. Rao *et al.* [20] devise a self-supervised attention-based gait encoding (AGE) model with multi-layer LSTM to encode gait features from unlabeled skeleton sequences, and then fine-tune the learned features with the supervision of labels for person re-ID. In [21], they further propose a locality-awareness approach (SGELA) that combines various pretext tasks (*e.g.*, reverse sequential reconstruction) and contrastive learning scheme to enhance self-supervised gait representation learning for the person re-ID task. The self-supervised work SM-SGE [23] utilizes a skeleton graph based reconstruction and inference mechanism to encode discriminative skeleton structure and motion features for the person re-ID task.

The most similar work to ours is [22]. Different from [22] that performs supervised skeleton representation for person re-ID, this work proposes the novel skeleton prototype contrastive learning (SPC) to achieve *unsupervised* skeleton-based person re-ID without using labels for more general settings. We for the first explore unified and generalizable multi-level (part-level, body-level, hyper-body-level) skeleton graphs to extend skeleton graph modeling to different skeleton

skeleton features that can re-identify different pedestrians under the unavailability of labels, which limits its application in many real-world scenarios.

To address the above challenges, this work for the first time proposes a generic Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation modeling (SPC-MGR) in Fig. 1 that can comprehensively model body structure and relations at various levels and mine discriminative features from *unlabeled* skeletons for person re-ID. Specifically, we first devise ***multi-level graphs*** to represent each 3D skeleton in a *unified coarse-to-fine* manner, so as to fully model body structure within skeletons. Then, to enable a comprehensive exploration of relations between different body components, we propose to model *structural-collaborative body relations* within skeletons from multi-level graphs. In particular, since each body component is highly correlated with its physically-connected components and may possess different *structural relations* (*e.g.*, motion correlations), we propose a ***multi-head structural relation layer*** (MSRL) to capture multiple relations between each body-component node and its neighbors within a graph, so as to aggregate key correlative features for effective node representations. Meanwhile, motivated by the fact that dynamic cooperation of body components in motion could carry unique patterns (*e.g.*, gait) [29], we propose a ***full-level collaborative relation layer*** (FCRL) to adaptively infer *collaborative relations* among motion-related components at both the *same-level* and *cross-level* in graphs. Furthermore, we exploit a multi-level graph feature fusion strategy to integrate features of different-level graphs via collaborative relations, which encourages the model to capture more graph structural semantics and discriminative skeleton features. Lastly, to mine effective features from *unlabeled* skeleton graph representations (referred as *skeleton instances*), we propose a ***skeleton prototype contrastive learning scheme*** *(SPC)*, which clusters correlative skeleton instances and contrasts their inherent similarity with the most representative skeleton features (referred as *skeleton prototypes*) to learn general discriminative skeleton representations in an unsupervised manner. By maximizing the similarity of skeleton instances to their corresponding prototypes and their dissimilarity to other prototypes, SPC encourages the model to capture more discriminative skeleton features and class-related semantics (*e.g.*, intra-class similarity) for person re-ID *without using any label*. The SPC is devised based on the proposed multi-level skeleton graph representations and structural-collaborative relation learning, and we experimentally and theoretically validate its effectiveness on unsupervised skeleton representation learning for person re-ID tasks.

Our main contributions are summarized as follows:

- We devise unified multi-level graphs to model 3D skeletons, and propose a novel Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation modeling (SPC-MGR) to learn an effective representation from unlabeled skeleton data for unsupervised person re-ID.
- We propose multi-head structural relation layer (MSRL) to capture relations of neighbor body components, and devise full-level collaborative relation layer (FCRL) to infer collaboration between different-level components, so as to learn more structural semantics and unique patterns.

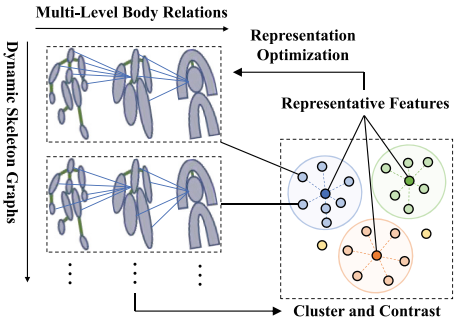


Fig. 1. Our approach constructs skeleton graphs to model multi-level body components and relations, and contrasts the clustered representative features to learn effective skeleton representations for person re-ID.

visual features such as human appearances, silhouettes and body textures from RGB or depth images to discriminate different individuals. Nevertheless, this kind of methods are often vulnerable to appearance, lighting and clothing variation in practice. Compared with RGB-based and depth-based methods, 3D skeleton-based models [18–23] exploit 3D coordinates of numerous key joints to characterize human body and motion, which enjoys smaller data size and better robustness to scale and view variation [24]. With these advantages, 3D skeleton data have drawn surging attention in the fields of person re-ID and gait recognition [20–23,25,26]. However, the way to model discriminative body and motion features with 3D skeleton data remains to be an open challenge.

To perform person re-ID via 3D skeletons, existing endeavors typically model skeleton features by two groups of methods. *Skeleton descriptor based methods* [18,19,27] manually extract certain anthropometric and geometric attributes of body from skeleton data. However, these hand-crafted methods usually require domain knowledge such as human anatomy [28], and cannot fully mine underlying features beyond human cognition. *Deep neural network based methods* [20,21,26] usually leverage convolutional neural networks (CNN) or long short-term memory (LSTM) to learn skeleton representations with sequences of raw body-joint positions or pose descriptors (*e.g.*, limb lengths). Nevertheless, these works rarely explore inherent relations between different body joints or components, which could ignore some valuable structural information of human body. Taking the human walking for example, neighbor body joints "foot" and "knee" have strong motion correlations, while they usually enjoy diverse degree of *collaboration* with limb-level components "leg" and "arm" during movement, which can be exploited to catch unique and recognizable patterns [29]. Other important flaws of this type of methods are label dependency and weak generalization ability. In practical terms, these methods usually require massive labeled data of pre-defined classes to either train the model from scratch [22,26] or fine-tune the pre-trained skeleton representations [20,21,23] to classify the known identities. As a result, they lack the flexibility to learn general and representative

Skeleton Prototype Contrastive Learning with Multi-level Graph Relation Modeling for Unsupervised Person Re-Identification

Haocong Rao[1,2] and Chunyan Miao[1,2(✉)]

[1] LILY Research Centre, Nanyang Technological University (NTU), Singapore, Singapore
{haocong001,ascymiao}@ntu.edu.sg
[2] School of Computer Science and Engineering, NTU, Singapore, Singapore
https://www.ntu.edu.sg/lily

Abstract. Person re-identification (re-ID) via 3D skeletons is an important emerging topic with many merits. Existing solutions rarely explore valuable body-component relations in skeletal structure or motion, and they typically lack the ability to learn general representations with unlabeled skeleton data for person re-ID. This paper proposes a generic *unsupervised* Skeleton Prototype Contrastive learning paradigm with Multi-level Graph Relation learning (SPC-MGR) to learn effective representations from *unlabeled* skeletons to perform person re-ID. Specifically, we first construct *unified multi-level skeleton graphs* to fully model body structure within skeletons. Then we propose a *multi-head structural relation layer* to comprehensively capture relations of physically-connected body-component nodes in graphs. A *full-level collaborative relation layer* is exploited to infer collaboration between motion-related body parts at various levels, so as to capture rich body features and recognizable walking patterns. Lastly, we propose a *skeleton prototype contrastive learning scheme* that clusters feature-correlative instances of unlabeled graph representations and contrasts their inherent similarity with representative skeleton features ("*skeleton prototypes*") to learn discriminative skeleton representations for person re-ID. Empirical evaluations show that SPC-MGR significantly outperforms several state-of-the-art skeleton-based methods under different scenarios.

Keywords: Skeleton Based Person Re-Identification · Unsupervised Representation Learning · Multi-Level Skeleton Graphs · Skeleton Prototype Contrastive Learning

1 Introduction

Person re-identification (re-ID) aims at identifying or matching a target pedestrian across different views or scenes, which plays an essential role in safety-critical applications including intelligent video surveillance, security authentication and human tracking [1–10]. Conventional studies [11–17] typically utilize

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 196–218, 2024.
https://doi.org/10.1007/978-3-031-65126-7_19

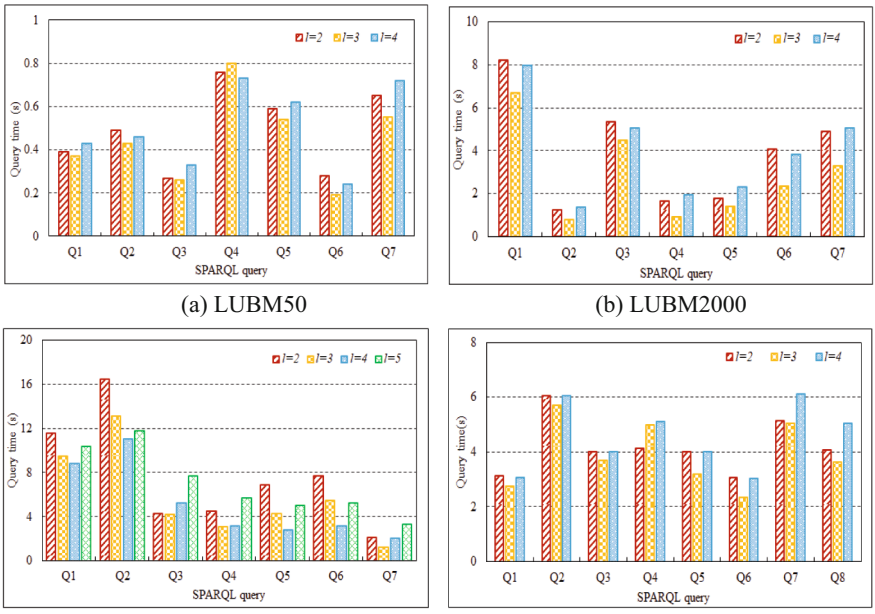
17. Udrea, O., Pugliese, A., Subrahmanian, V.S.: GRIN: a graph based RDF index. In: 22th AAAI Conference on Artificial Intelligence, British Columbia, Canada, pp. 1465–1470(2007)
18. Zou, L., Özsu, M.T., Chen, L., et al.: GStore: a graph-based SPARQL query engine. VLDB J. **23**(4), 565–590 (2014)
19. Wang, D., Zou, L., Feng, Y., et al.: S-store: an engine for large RDF graph integrating spatial information. In: 18th International Conference on Database Systems for Advanced Applications, Wuhan, China, pp. 31–47 (2013)
20. Tran, T., Ladwig, G.: Structure index for RDF data. In: Proceeding of the Workshop on Semantic Data Management (2010)
21. He, H., Wang, H., Yang, J., et al: BLINKS: ranked keyword searches on graphs. In: The ACM SIGMOD International Proceedings on Management of Data,Beijing, China, pp. 305–316 (2007)
22. Kim, K., Moon, B., Kim, H.J.: R3F: RDF triple filtering method for efficient SPARQL query processing. World Wide Web-Internet Web Inf. Syst. **18**(2), 317–357 (2015)
23. Brisaboa, N.R., Ladra, S., Navarro, G.: Compact representation of Web graphs with extended functionality. Inf. Syst. **39**(1), 152–174 (2014)
24. Schmidt, M., Guo, Y., Pan, Z., Heflin, J.: LUBM: a benchmark for OWL knowledge base systems. Web Semant. Sci. Serv. Agents World Wide Web **3**(2), 158–182 (2005)
25. Hornung, T., Lausen, G., et al: SP2Bench: a SPARQL performance benchmark.In: 25th International Proceedings on Data Engineering, Shanghai, China, pp. 222–233 (2009)
26. Uniprot RDF. http://dev.isb-sib.ch/projects/uniprot-rdf/

parallel queries. Of course, distributed query systems involve communication and load balancing issues, which will also be our future research direction.

Acknowledgments. This work is partially supported by the Scientific research project of The Educational Department of Liaoning Provincial under Grant LJ2020016 and Research Institute Project of Bohai University under Grant XK202134-3.

References

1. Ning, Z., Huang, J., Wang, X.: Vehicular fog computing: enabling real-time traffic management for smart cities. IEEE Wirel. Commun. **26**(1), 87–93 (2019)
2. Pirrò, G.: Building relatedness explanations from knowledge graphs. Semant. Web **10**(6), 963–990 (2019)
3. Ning, Z., Kwok, R., Zhang, K., et al.: Joint computing and caching in 5G-envisioned Internet of Vehicles: a deep reinforcement learning-based traffic control system. IEEE Trans. Intell. Transp. Syst. **22**(8), 5201–5212 (2020)
4. Feng, J., Meng, C., Song, J., et al.: SPARQL query parallel processing: a survey. In: IEEE International Conference on Big Data, Boston, USA, pp. 444–451 (2017)
5. Wang, X., Ning, Z., et al.: Offloading in Internet of Vehicles: a fog-enabled real-time traffic management system. IEEE Trans. Ind. Inform. **14**(10), 4568–4578 (2018)
6. Neumann, T., Weikum, G.: The RDF-3X engine for scalable management of RDF data. VLDB J. **19**(1), 91–113 (2010). https://doi.org/10.1007/s00778-009-0165-y
7. Weiss, C., Karras, P., Bernstein, A.: Hexastore: sextuple indexing for semantic web data management. In: 34th International Conference on VLDB, Auckland, New Zealand, pp. 1008–1019 (2008)
8. Mulay, K., Kumar, P.S.: SPOVC: a scalable RDF store using horizontal partitioning and column oriented DBMS. In: 4th International Workshop on Semantic Web Information Management, Scottsdale, USA, pp. 1–8 (2012)
9. Atre, M., Chaoji, V., Zaki, M.J., et al.: Matrix "Bit" loaded: a scalable lightweight join query processor for RDF data. In: 19th International Conference on World Wide Web, Raleigh, USA, pp. 41–50 (2010)
10. Matono, A., Pahlevi, S.M., Kojima, I.: RDFCube: a P2P-based three-dimensional index for structural joins on distributed triple stores. In: The International Conference on Databases, Information System, and Peer-to-Peer Computing, Seoul, Korea, pp. 323–330 (2005)
11. Yuan, P., Liu, P., Wu, B., et al.: TripleBit: a fast and compact system for large scale RDF data. In: 39th International Conference on VLDB, Trento, Italy, pp. 517–528 (2013)
12. Harris, S., Gibbins, N.: 3store: efficient bulk RDF storage. In: 1st International Workshop on Practical and Scalable Semantic Systems,Florida, USA, pp. 1–15 (2003)
13. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: a generic architecture for storing and querying RDF and RDF schema. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 54–68. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-48005-6_7
14. Carroll, J.J., Dickinson, I., Dollin, C., et al.: Jena: implementing the semantic web recommendations. In: 13th International World Wide Web Conference on Alternate Track Papers & Posters,New York, USA, pp. 74–83 (2004)
15. Abadi, D.J., Marcus, A., Madden, S.R., et al.: SW-store: a vertically partitioned DBMS for Semantic Web data management. VLDB J. **18**(2), 385–406 (2009)
16. Ning, Z., Xia, F., Ullah, N., Kong, X., Xiping, Hu.: Vehicular social networks: enabling smart mobility. IEEE Commun. Mag. **55**(5), 16–55 (2017). https://doi.org/10.1109/MCOM.2017.1600263



(a) LUBM50 (b) LUBM2000

(c) Uniprot (d) SP^{2}Bench

Fig. 4. Comparision of query process time on different values of parameter *l*

5 Conclusions

Aiming at the frequent self joins in triple based retrieval and the semantic association characteristics reflected by chain structure information in SPARQL complex query. This paper proposed a bit index structure based on path (PathBit) for large scale RDF Graph. PathBit created predicate path tree (IPT) to filter complete path sets associated with SPARQL query and designed a k^2-tree index (k^{2}TIP) according to the hierarchy of each predicate path tree. k^{2}TIP realized fast association matching of known predicate path triples. Meanwhile, the compression mechanism is used to implement the compressed storage and retrieval algorithm of triples. In addition, two auxiliary indexes: SP and OP are added to assist predicate path retrieval. In the experiment, we compare PathBit with the three existing index storage schemes. The experimental results show that PathBit is very effective for complex queries, especially for queries with long paths. And with the expansion of data scale, PathBit has higher retrieval advantages. At the same time, the storage space of the compressed storage method used in this paper is 0.96 times less than that of RDF-3X in four datasets under the parameter of combined prefix, and has certain advantages over Triplebit.

In addition, with distributed data processing become mainstream, distributed indexing and querying have become research hotspots. The index structure of PathBit can be divided into several sub trees, and the leaf nodes of each sub tree are related to a complete set of paths. Allocating these sub trees to various computing nodes can achieve

4.4 Parameter Analysis

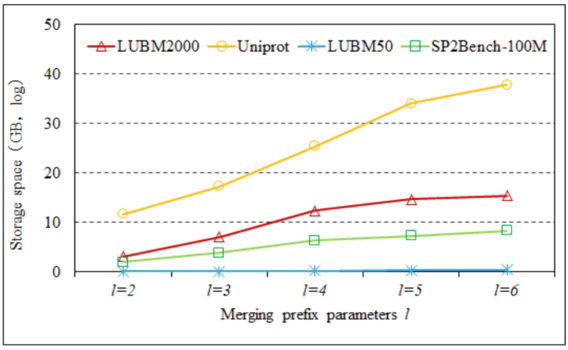


Fig. 3. Comparison of storage space on different values of parameter l

In the process of predicate path merging, the length of common prefix edge is controlled by parameter l. When the length of the common prefix is greater than or equal to l, the edges with the common prefix are merged. This part will test the influence of the parameter l on the storage space.

The range value of l is from 2 to 6. Figure 3 shows that the storage space of all datasets increases with the increase of l. The reason is that the larger the value of l, the smaller the probability of having a common prefix, and the fewer replica nodes that can be merged. In addition, Fig. 3 also shows that with the gradual increase of l value, the change of storage space will be smaller and smaller. When l is set to 4 or 5, the storage space tends to be stable for LUBM and Sp^{2}bench datasets. However, for UniProt, l varies from 5 to 6, because the predicates in UniProt are larger than the other two datasets, which makes the length of common prefix between paths longer.

Figure 4 shows a comparison of query performance. The experimental results show that l has an optimal value, but this value is not directly proportional to the value of l. As shown in Fig. 4, the optimal value of l is 3 for LUBM and Sp^{2}bench datasets, and 4 for UniProt dataset. There are two main reasons. First, when l value is too small, a large number of predicate path are merged, resulting in more candidate paths in the path template matching, which affects the final path matching efficiency. On the contrary, when the value of l is too large, the candidate set becomes smaller and the number of copies increases, which also reduces the query efficiency. Considering the storage space and query performance, the query performance is the best when l takes the storage space to be stable.

The size of Unirpot dataset is 700 million. When the merging common prefix parameter l is set to 3 and the current hardware environment is used to execute the query, Bitmat cannot get the query result. Therefore, Fig. 2(c) only lists the query time comparison of RDF-3X, Triplebit and PathBit. Similarly, after PathBit decomposes the query into query paths, many join operations between triples are decomposed into each path set to complete separately, and the intermediate result set involved in join becomes smaller, especially Q1, Q2, Q5, Q6 and Q7 contain long paths. Combined with the auxiliary indexes, the query performance is superior to Triplebit and RDF-3X. However Q3 and Q4 are star queries, so the overall performance is not as good as long-path retrieval.

The same index is also used to execute queries on SP^2Bench dataset. As with LUBM dataset, the merging common prefix parameter l is also set to 3. Figure 2(d) shows the execution time of each query. Since most of SP^2Bench standard data queries are star structured and query design pays more attention to the use of query operators, the overall query efficiency takes less time, but the filtering and merging operations of the results take a long time. Figure 2(d) depicts the time taken to execute a basic query. Among them, Q1, Q2, Q3 and Q4 are star queries, which are comparable to Triplebit, but better than RDF-3X. However, Q5 and Q7 contain the long path queries, especially Q7, so the query efficiency is significantly improved.

4.3 Comparison of Storage Space

In this part, we compare the storage space of PathBit, RDF-3X and Triplebit. Here, the storage space refers to the space consumed by storing datasets and indexes. Since Bitmat does not contain dictionary tools, the comparison results don't include Bitmat. The merging common prefix parameter l is also set to 3. Table 4 lists the space consumed of different datasets. It can be seen that the storage space of PathBit on all datasets is lower than the other three indexes. As explained earlier, RDF-3X needs to create 6 cluster indexes and 9 clustered indexes. As we all know, the high efficiency of RDF-3X is at the cost of storage space. The dictionary tool used by PathBit is the same as Triplebit. Due to the small number of predicates in LUBM and SP^2Bench and the high merging rate of predicate path, the number of copies is greatly reduced. Therefore, PathBit is better than Triplebit on these two datasets. But the storage space on UniProt dataset is higher than Triplebit.

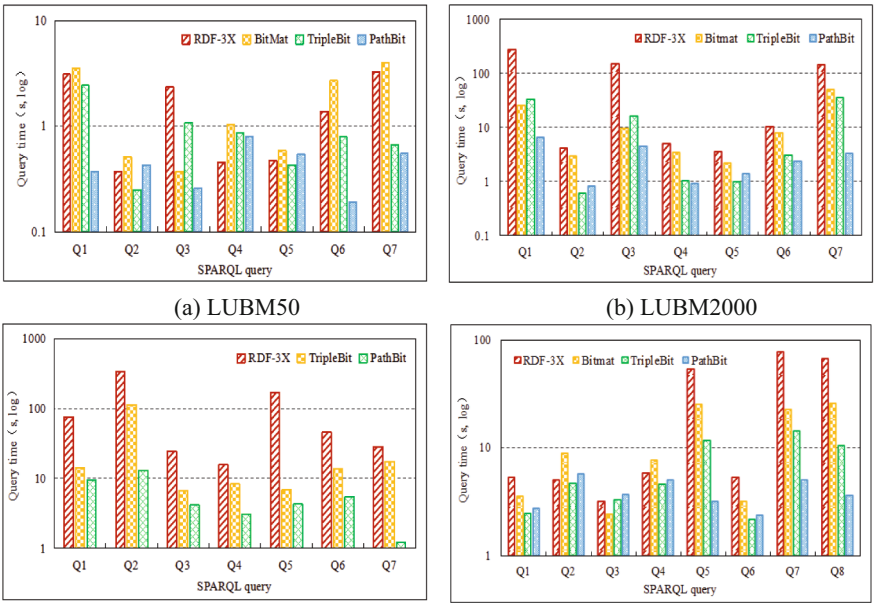
Table 4. Comparision of storage space (GB)

	RDF-3X	TripleBit	PathBit
LUBM50	0.35	0.28	0.19
LUBM2000	13.95	8.74	7.11
Uniprot	33.89	15.19	17.28
SP^2Bench	7.28	4.17	3.88

of query efficiency. The intermediate result in this paper refers to the number of triples matched with the query and the data loaded into memory during the query. Because of the compression method and the direct search in the compressed form, PathBit loads more query data in the same memory and reduce the I/O cost.

It also shows that PathBit is very effective in retrieving large data sets. When the size of LUBM increases from 50 to 2000, the minimum change of query time on RDF-3X is 7.5 times, and the maximum change is 90.83 times, especially for complex queries Q1 and Q3. However, the maximum change of PathBit was only 18.11 times. The reason is that RDF-3X query needs to load more indexes into memory, and at the same time, it also needs to decompress. Therefore, the I/O is larger. Bitmat and Triplebit are both based on triple mode.

In the face of complex queries, they need to join and merge triple more times, so the query performance is lower than that of PathBit. Q2, Q4 and Q5 are star queries. The semantic relevance of predicate path information obtained by star structure is relatively low, but the subject set meeting the conditions is obtained by auxiliary index SP. Combined with subject set and hierarchical edge index, a large number of unrelated triples can be filtered, and the scale of merging results can be reduced. Therefore, the execution efficiency on LUBM 2000 dataset is still better than RDF-3X and Bitmat, which is equivalent to Triplebit. Because the LUBM 50 dataset is small, the index can load memory at once, so RDF-3X retrieval is the highest.



(a) LUBM50 (b) LUBM2000

(c) Uniprot (d) SP²Bench

Fig. 2. Comparision of query performance

Table 2. Connection types of triple pattern2

Category	Type
1	$s\ ?p_1\ ?x \bowtie ?x\ ?p_2\ o$
2	$?s\ ?p_1\ ?x \bowtie ?x\ ?p_2\ o$
3	$?s\ ?p_1\ ?x \bowtie ?x\ ?p_2\ ?o$

4 Experiments

In this section, PathBit indexing scheme is tested on synthetic and real datasets.

4.1 Datasets and Setting

Table 3. Test datasets

Data set	Vertex	Triple	Predicate
LUBM50	1,706,230	6,888, 642	18
LUBM2000	66,059, 204	276, 345,040	18
SP2Bench	56,125,032	113,246,165	22
Uniprot	139,942,781	687,025,165	84

In the experiment, two synthetic datasets LUBM and Sp^2Bench were selected. The LUBM features a university domain, and the SP^2Bench dataset features a DBLP domain [24, 25]. In our experiments, we also use a protein dataset Uniprot [26] (Table 3).

PathBit index is written in C++ and compiled with GCC. We select the optimization level of O2. The experiment runs on a server with Intel Xeon 2.00GHz processor and 20GB memory. Considering the influence of warm cache on experimental error, each query is executed five times, and the arithmetic average is taken as the final experimental result.

4.2 Comparison of Query Performance

In the experiment, LUBM data set generates 81 predicate paths. If the merging common prefix parameter l is set to 3, we obtain 26 predicate path trees. Figure 2(a) and (b) show the query execution time of SPARQL. The query time of Q1, Q3, Q6 and Q7 are better than the other indexes. These four queries have longer join paths than the star queries Q2, Q4 and Q5. Using the path association information to search can filter a large number of unrelated triples and narrow the retrieval range. Hierarchical path index decomposes the connection between triples into smaller ones, which reduces the connection size of triples and improves the matching efficiency. Intermediate results are also an affecting factor

According to Theorem 4, given a SPARQL query G_q, G_q is decomposed into a complete path set $QCPath = \{R_1^q, R_2^q, \ldots R_n^q\}$. If G_q is a subgraph of G, then $\forall R_i^q \in QCPath$, there must exist at least one complete path R_i, satisfying R_i^q is a subpath of R_i. Therefore, a SPARQL query need to be decomposed into multiple search paths from the source vertex to the sink vertex. The decomposition principle is to follow the full coverage of vertices and edges. That is, starting from any source vertex, if the decomposed full path already contains all the edges in the query, the decomposition ends. Because the complete path decomposition already includes all possible complete paths, there must be a complete path corresponding to it.

According to whether the predicate path contains a constant, the decomposed search path can be divided into two categories: constant predicate path and variable predicate path. Constant predicate refers to the path containing one or more known predicates, while variable predicate refers to the path in which all predicates are unknown.

The analysis result shows that most of predicate paths of SPARQL queries are constant predicate path. For constant predicate paths, the retrieval is performed on the IPT index to obtain the candidate complete paths containing known predicates. Then according to the connection relationship of adjacent predicates, they are divided into six types as shown in Table 1. When performing the retrieval, the type of adjacent predicate is judged from the source vertex in turn, and is executed in the order from low to high.

Table 1. Connection Types of triple pattern1

Category	Type
1	$s\,p_1\,?x \bowtie ?x\,p_2\,o$
2	$?s\,p_1\,?x \bowtie ?x\,p_2\,o$
3	$?s\,p_1\,?x \bowtie ?x\,p_2\,?o$
4	$s\,p_1\,?x \bowtie ?x\,?p_2\,o$
5	$?s\,?p_1\,?x \bowtie ?x\,p_2\,o$
6	$?s\,?p_1\,?x \bowtie ?x\,p_2\,?o$

When the search path is a variable predicate path, the adjacent predicate connections can be divided into three types as shown in Table 2. Since k^2TIP is only applicable to the case that there is a constant predicate in the retrieval path. For the variable predicate path, not only IPT is invalid, but k^2TIP is also invalid. In order to ensure the validity of the index, we design two auxiliary indexes, namely SP and OP, to solve this problem. SP and OP store all predicates corresponding to each subject or object, respectively. SP and OP indexes adopt the compressed representation and retrieval method proposed in Triplebit, which will not be explained in detail here.

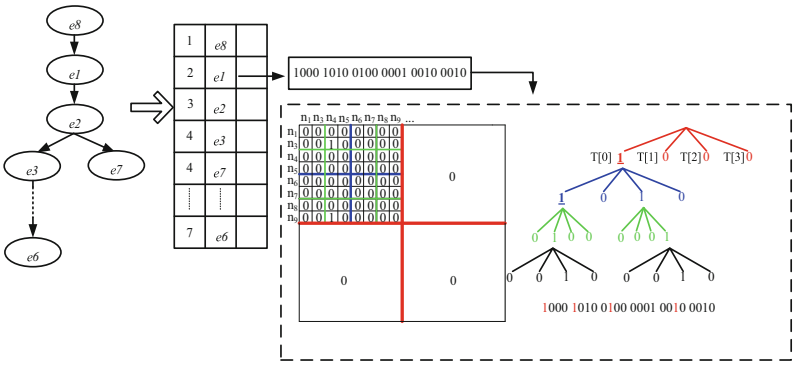


Fig. 1. Example of k^2TIP

node of the root node in k^2 tree. If the bit element of the sub matrix contains 1, then the corresponding sub node of k^2 tree is 1, otherwise the sub node is 0. After the first level node is created, the matrix corresponding to the node with the value of 1 will continue to be divided in the same way until the sub matrix is 0 or there are only k^2 bits in the sub matrix.

The generated k^2-tree connects nodes corresponding to 0 or 1 from top to bottom and left to right to form a compressed bit string. Figure 1 describes the generation process of bit matrix, k^2-tree and compressed bit string of a predicate path tree. For a triple set of a predicate, it is usually necessary to search all the objects corresponding to a known subject, or to search all the subjects corresponding to the known subject. As shown in Fig. 2 shows all objects associated with the subject n_3, where the id of n_3 is 2 and the value of k is 2. In order to get the all objects associated with n_3, we have to find all the columns with cell 1 in the row of n_3 in the matrix. The specific steps are as follows:

Step 1: starting from the root node of k^2-tree, set its position *pos* as 0, and its four sub nodes corresponding to the four sub matrices respectively. The sub matrix corresponding to the first two sub nodes intersect with the row of n_3, while the other two child nodes has no association with n_3. So we only need to consider the first two sub nodes. Set the two sub matrices to T[0] and T[1]. T[1] sub matrix is 0, which means that there is no object associated with n_3 in T[1]. In order to get the number of the column in which the object associated with n_3, it is necessary to record the starting position of the corresponding column of the incidence sub matrix when locating the incidence sub matrix. For example, the starting position of T[0] sub matrix is 0.
Step 2: in the compressed bit string, the starting position *pos* of sub node corresponding to T[0] is 4. The corresponding bit of four sub matrices of T[0] is '1010'. The id of n_3 indicates that the sub matrix associated with it is T[0][0] and T[0][1], where T[0][0] is 1 and T[0][1] is 0. And the starting position of T[0][0] sub matrix is 0.
Step 3: according to this method, continue to search down until the leaf node. If the leaf node is 1 and satisfies the association with n_3, then the column corresponding to the node is the object associated with n_3.

bit of the bit string corresponds to a unique predicate. Supposing $E(ppt_i)$ is the predicate set of a predicate path tree, where ppt_i is the ith predicate path tree in *PPtree*. If $\exists pre \in ppt_i$, set the *eId* (*pre*) bit of the bit string to 1, where *eId* (*pre*) represents the id of the predicate *pre*.

The establishment of IPT is based on the bottom-up process. Each leaf node of IPT corresponds to a predicate path tree, and each path template tree corresponds to a full path set. Non-leaf nodes of IPT are obtained by performing a logical 'OR' operation on their sons.

3.3 Retrieval of Predicate Path Tree

When retrieving the query predicate path tree on the IPT index tree, we encode the query path tree in the same way according to Definition 8. If the query predicate path tree includes predicate variables, the code of predicate variable is set to 0. Then use the top-down method to search for the match paths in IPT index tree. If the query predicate path tree and the node of IPT meet the matching principle of Definition 9, the search continues, otherwise the subtrees corresponding to the unmatched node on IPT index tree are pruned.

Definition 9 (Matching Principle). Given a bit string of predicate path tree bit string *ppt** and a query path bit string *qtt**, if *ptt** matches *qtt**, if and only if the logical 'and' operations satisfy AND(*ptt**, *qtt**) = *qtt**.

3.4 Match of Complete Path

Retrieved results in IPT are a candidate set of complete path. In order to get the final query results, it is necessary to accurately match the candidate path set. In this section, we will create a k^2-tree index (k^2TIP) for the corresponding complete path collection according to the hierarchy of each predicate path tree. The k^2TIP adopts two stage compression modes. Figure 1 shows the k^2-tree index structure. The k^2TIP index contains each edge in the predicate path template tree and its corresponding hierarchical ID information, and each edge points to a storage area, which is used to store the triple set associated with the edge in the whole path. Since all triples of this set have a common predicate, we use k^2-tree [23] structure to store triples for each triple set, and compress triples on this basis.

Each predicate in the predicate path tree corresponds to a triple set, and these triples share the same predicate. In order to reduce the storage space, we compress each triple set, and the compression method adopts k^2-tree.

k^2-tree first uses a two-dimensional bit matrix to establish the corresponding relationship between subject and object. If there is a corresponding relationship between the subject and the object, the corresponding bit is set to 1, otherwise it is 0. A large number of subjects and objects are not related, so the bit matrix is a sparse matrix. Therefore, we divide the bit matrix into k^2 sub matrices, and each sub matrix corresponds to a sub

Proof: Suppose there is no complete path $Ri \in CPath$, satisfying R_i^q is a subpath of R_i. (1) If the source vertex v_0 and sink vertex v_m of the complete query path R_i^q in G_q is also the source and sink vertices of G, then this will conflict with the hypothesis, because there must be a complete path between v_0 and v_m. (2) If the source vertex v_0 and sink vertex v_m of R_i^q in G_q is not the source and sink vertices of G, then there must be a source vertex v_s, which v_s to v_0 is reachable and there is also a sink vertex v_e, which v_m to v_e is also reachable. That is to say, there is at least one complete path from v_s to v_e and v_0 to v_m is a subpath of the whole path, which contradicts the hypothesis. To sum up, the hypothesis does not hold, that is, there is at least one complete path $R_i \in CPath$, satisfying R_i^q is a subpath of R_i.

Definition 5 (Predicate Path). Given an RDF graph G, according to the Definition 4, G is decomposed into a set of complete paths. We use $CPath = \{R_1, R_2, \ldots, R_m\}$ to represent it. For any path R, Extract the edge information of the path to construct a summary path $E(R_i) = \{e_1, e_2, \ldots, e_m\}$, then $E(R_i)$ is called predicate path.

Definition 6 (Isomorphism Path). If the complete paths have the same predicate path, they are called to be isomorphic paths.

Obviously, after decomposing the RDF graph into the set of complete path, a large number of vertices and edges repeatedly appear in different complete paths, which puts a lot of pressure on data storage. However, many complete paths have similar predicate path, that is, many vertices information is the same. If these complete paths are divided into the same class, the number of copies of vertices will be reduced. So we define the following two conditions to merge predicate path.

Definition 7 (Predicate Path Tree). If two or more predicate paths meet the following conditions: (i) Two or more predicate paths have a common prefix and the length of the edge of the common prefix is greater than or equal to a threshold. (ii) A predicate path is the suffix of another predicate path. We will merge these paths into one predicate path tree. The predicate path tree is denoted as *PPtree.*

3.2 Index of Predicate Path Tree

Each predicate path tree corresponds to a complete path set. For a SPARQL query, it is decomposed into several query paths in the same way. We can obtain the complete path set corresponding to each query path by retrieving the predicate path tree. In order to quickly locate the complete path set of the query paths, we create an index based on predicate path tree (IPT). The establishment process of IPT mainly include three steps: the first is to code the predicate path tree, the second is the construction process of IPT and the last is how to retrieve IPT.

Definition 8 (Encoding Predicate Path Tree). Assign a unique id to each predicate in L in order. Obviously, the maximum id is the number of elements in L, which is represented by ℓ. The encoding of predicate path tree is a bit string of length ℓ, and each

filtering operations are used to filter out irrelevant data in the input triples. TripleBit [11] vertically divides the triple matrix based on predicates, and sorts triples with the same predicate in the order of subject or object. During the query process, two index structures was introduced to minimize the cost of index selection. In summary, path, compression and index tree are very effective techniques for improving query efficiency and reducing storage space.

3 PathBit

PathBit includes an index based on predicate path tree (IPT) and a k^2-tree index (k^2TIP) according to the hierarchy of each predicate path tree. IPT is in charge of the filter of complete path set, which related to the retrieval path. k^2TIP according to the hierarchy of each predicate path tree to realize fast association matching of known predicate path triples. Meanwhile, the compression mechanism is used to implement the compressed storage and retrieval algorithm of triples. In addition, two auxiliary indexes: SP and OP are added to assist predicate path retrieval.

3.1 Complete Predicate Path

An RDF database is a set of RDF triples, we use $T = \{t \mid t \in S \times P \times O\}$ to describe the dataset, where S, P, O are the set of subjects, predicates and objects, respectively.

Definition 1 (Path). Given an RDF $G = (V, E, L)$, a path is a set of ordered vertices, denoted by $R = (v_0 v_1 v_2 \ldots v_m)$, $\forall k \in [0, m-1]$, $<v_k, v_{k+1}> \in E$.

Definition 2 (Complete Path). For any path R in RDF graph, if v_0 is a source vertex and v_m is a sink vertex, we say that R is a complete path. We use $CPath = \{R_1, R_2, \ldots, R_m\}$ *to* denote a set of RDF complete paths.

Theorem 3. Given an RDF G, $\forall v \in V$ and $e(u, v) \in E$ must belong to at least one complete path.

Proof: (1) Assuming SV is the set of source vertices. For any vertex v in graph G, there are two states. The first is $v \in SV$. If $v \in SV$, because any complete path starts from a source vertex, v must exist in a complete path. The second is $v \notin SV$. If $v \notin SV$, then there must be a source vertex s, so that s to v can be reached, that is, the vertex v belongs to a complete path whose source vertex is s. If s doesn't exist, then v becomes the source vertex, which conflicts with the condition. Therefore, for any vertex v in set V must belong to at least one complete path. (2) For any edge $e(u, v) \in E$ in graph G, if it does not belong to any complete path, then the two vertices u or v do not exist in any complete path, which is in contradiction with that any vertex $v \in V$ belongs to at least one complete path. Therefore, any edge $e(u, v) \in E$ belongs to at least one full path.

Theorem 4. Given a SPARQL query G_q, according to the Definition 4, G_q is decomposed into a set of complete query paths. We use $QCPath = \{R_1^q, R_2^q, \ldots R_n^q\}$ to represent it. If G_q is a subgraph of G, then $\forall R_i^q \in QCPath$, there must exist at least one complete path R_i, satisfying R_i^q is a subpath of R_i.

2 Related Work

In order to improve the retrieval efficiency, researchers have conducted extensive research on the storage and index of RDF. This paper analyzes the current research status from three different perspectives.

RDF storage and index technology based on relationships utilizes relational database query technology to convert SPARQL queries into SQL to realize data retrieval. 3-Store [12] and Sesame [13] all use triple tables. Due to all data exists in a large table, SPARQL queries are easy to result in many self joins and decrease the query efficiency. Jena2 [14] uses an attribute table, which greatly reduces self joins and merge operations. But not all objects have the same properties, which leads to a large number of empty values. In addition, a large number of multi-valued attributes can also generate more multi-valued dependencies. Therefore, the attribute table is not a universal storage model. SW-store [15] decomposes triples based on the predicate, storing triples with the same predicate in the same table. For the two columns table, a subject based clustered index can be created to achieve rapid subject localization. This scheme not only reduces the merging operation of the same predicate, but also avoids the control problems caused by the attribute table.

Triple index scheme is a combination and permutation of S, P and O. RDF-3X [6] stores all permutations and combinations of subject, predicate and object on a B+ tree, respectively. Moreover, RDF-3X also combines two or single elements to directly form a clustered index. Similar to RDF-3X, Hexastore [7] also establishes six indexes based on the triple table. The difference is that in the establishment process of index, Hexastore considers the order relationship between the subject, predicate and object. Meanwhile, Hexastore will reduce the redundancy of memory by sharing index lists. SPOVC [8] creates five index types based on subject, predicate, object, object data types, and triple classes. Each index type was horizontally segmented according to certain rules, which is effective for the query of range or rule expressions.

Bitmat [9] and RDFcube [10] map S, P and O into a three-dimensional space to form a three-dimensional matrix. Each element in the matrix corresponds to a triple. Bitmat is a memory based bit matrix primarily used to handle concatenation operations in triple pattern. Although these two index types utilize bit technology to achieve high compression of triples, they face large-scale data, especially Bitmat, which makes it difficult to load the index into memory at once.

RDF itself is a directed graph, so SPARQL query can be seen as a sub graph matching problem. GRIN [17] indexes RDF graphs with a balanced binary tree. By utilizing the distance conditions, it can quickly filter the data that does not meet the criteria. But GRIN index has poor scalability. Zou et al. [18, 19] proposed VS-tree and VS*-tree index to handle precise and wildcard SPARQL queries. PIG [20] (Parameterized Index Graph) index corresponds to a set of vertices with similar or identical neighborhood structures in the original data graph. PIG first retrieve edges that are homomorphic to the edges in the query graph to form a set of candidate edges, and then perform join operations in the set of candidate edges. He et al. [21] proposed a two-layer index scheme (BLINKS) for searching the top-k keywords on a graph, which only supports searching on node labeled directed graphs. In order to reduce redundant intermediate results, RP-index [22] creates a path based index to index the RDF graph in-edge. During the executive process,

The rapid increase of RDF data brings great challenges to traditional data storage, index and query. The triple table, vertical partition and attribute table use an alternative relational storage mode and mature management mechanism of relational database to accelerate data retrieval. However, these relational data models could not fully reflect the logical structure of RDF data [5].

Some native storage systems, such as RDF-3X [6], Hexastore [7] and SPOVC [8], store multiple copies of data according to different combinations of subject, predicate and object to assist in generating better query plan. Although the query efficiency is improved, these systems are at the expense of storage space. Bitmat [9] and RDFcube [10] use three-dimensional matrix to store triples, and divide the three-dimensional matrix into two-dimensional matrices along a certain dimension. For each two-dimensional matrix, D-gap compression method is used for row compression storage. However, in the face of large-scale data, it is difficult for Bitmat to load all indexes into memory at one time. Triplebit [11, 12] reduces the storage scale of RDF-3X, and only stores two combinations of subject and object (SO) and object and subject (OS) based on predicate. At the same time, Triplebit establishes the corresponding index according to the predicate and realizes the compressed storage. For a SPARQL query, Triplebit generates the corresponding query plan according to certain heuristic rules, and dynamically modifies the query plan to reduce the intermediate results.

All these storage systems view triple mode as retrieval unit to realize data retrieval. Through certain query plan and optimization technology, these storage systems reduce intermediate results, and realize fast connection of intermediate results. These indexes are based on triples and do not consider the structure and semantics of RDF graphs. As mentioned above, triple patterns in SPARQL queries have certain connection relations, which not only reflect the structural information of RDF graphs, but also reflect certain semantic relations. This paper proposes a bit index structure based on path (PathBit) for large scale RDF Graph. The major contributions include:

(1) Taking the complete path from source to sink point in RDF graph as the structure object, we create the bit index based on predicate path tree to realize the retrieval and filtering mechanism, and reduce the connection scale of intermediate results in triple pattern matching;
(2) For each complete path tree, we will create a k^2-tree index (k^2TIP) according to the hierarchy of each predicate path tree to realize fast association matching of known predicate path triples;
(3) We use k^2-tree compression mechanism to implement the compressed storage and retrieval algorithm of triples. At the same time, two auxiliary indexes: SP and OP are added to assist predicate path retrieval.

The other parts of the paper are summarized as follows: part two introduces the relate works; part three describes the design scheme of PathBit in detail; finally, experiments verify the performance of PathBit and draw the conclusion.

PathBit: A Bit Index Based on Path for Large-Scale Knowledge Graph

Yonglin Leng(✉), Peiyi Qu, Ying Guo, and Chaoliang Xi

College of Information Science and Technology, Bohai University, Jinzhou 121000, China
lengyonglin@qq.com

Abstract. As the latest achievement of symbolism, knowledge graph is an important cornerstone of artificial intelligence. In order to better manage the knowledge graph, RDF triples have been used to represent knowledge graph. The rapid growth of data brings great challenges to knowledge graph storage and quick retrieval. Among them, self joins, high storage cost and intermediate results are the main problems. In this paper, we propose a bit index structure based on path (PathBit) for large scale knowledge graph. PathBit includes an index based on predicate path tree (IPT) and a k^2-tree index (k^2TIP) according to the hierarchy of each predicate path tree. IPT is in charge of the filter of complete path set. k^2TIP according to the hierarchy of each predicate path tree to realize fast association matching of known predicate path triples. Meanwhile, the compression mechanism is used to implement the compressed storage and retrieval algorithm of triples. In addition, two auxiliary indexes: SP and OP are added to assist predicate path retrieval. Finally, we conduct a series of experiments on two representative datasets and compare the results with RDF-3X, Bitmat and TripleBit. Results indicate that PathBit can achieve better response time on complex queries and has greater advantages in storage space compared with RDF-3X and Bitmat.

Keywords: Knowledge Graph · Index · Predicate Path · Compressed storage

1 Introduction

As the supporting foundation of AI, knowledge graph shows more and more value in semantic search, intelligent question answering, data analysis, natural language processing, vision understanding and IoT. In order to better manage the knowledge graph, RDF triples have been used to represent knowledge graph. The rapid growth of RDF data also bring great challenges to query. SPARQL is the most widely used query language in RDF data query [1, 2]. A SPARQL query includes many triple patterns, which can also be described as a directed query graph. The query graph generally consists of four basic sub graphs: star, chain, ring and tree topology [3, 4]. The basic sub graph has a lot of connections. These connection relations can be divided into chain and star relations. The chain relation refers to the subject of triple pattern is the object of another triple pattern. Star relation refers to a group of triple patterns with the same subject or object. In these basic sub graphs, the chain relation is an important structure in SPARQL. Because the ring and tree query all contain chain structure.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 181–195, 2024.
https://doi.org/10.1007/978-3-031-65126-7_18

Reliability and Scalability

2. Mahmoud, M.S., Hamdan, M.M., Baroudi, U.A.: Modeling and control of cyber-physical systems subject to cyber attacks: a survey of recent advances and challenges. Neurocomputing **338**, 101–115 (2019)
3. Deng, Z., Xiong, C., Cai, M.: An autonomous transportation system architecture mapping relation generation method based on text analysis. IEEE Trans. Comput. Soc. Syst. **9**(6), 1768–1776 (2022)
4. Zhou, Z., Cai, M., Xiong, C., et al.: Construction of autonomous transportation system architecture based on system engineering methodology. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), pp. 3348–3353. IEEE (2022)
5. You, L., He, J., Wang, W., et al.: Autonomous transportation systems and services enabled by the next-generation network. IEEE Netw. **36**(3), 66–72 (2022)
6. Graja, I., Kallel, S., Guermouche, N., et al.: A comprehensive survey on modeling of cyber-physical systems. Concurr. Comput. Pract. Experience **32**(15), e4850 (2020)
7. Tantawy, A., Abdelwahed, S., Erradi, A., et al.: Model-based risk assessment for cyber physical systems security. Comput. Secur. **96**, 101864 (2020)
8. Schranz, M., Di Caro, G.A., Schmickl, T., et al.: Swarm intelligence and cyber-physical systems
9. Deka, L., Khan, S.M., Chowdhury, M., et al.: Transportation cyber-physical system and its importance for future mobility. Transp. Cyber-Phys. Syst. 1–20 (2018)
10. Hussain, M.D.M., Beg, M.M.S.: Using vehicles as fog infrastructures for transportation cyber-physical systems (T-CPS): fog computing for vehicular networks. Int. J. Softw. Sci. Comput. Intell. (IJSSCI) **11**(1), 47–69 (2019)
11. Lin, J., Yu, W., Zhang, N., et al.: Data integrity attacks against dynamic route guidance in transportation-based cyber-physical systems: modeling, analysis, and defense. IEEE Trans. Veh. Technol. **67**(9), 8738–8753 (2018)
12. Hou, Y., Zhao, Y., Wagh, A., et al.: Simulation-based testing and evaluation tools for transportation cyber–physical systems. IEEE Trans. Veh. Technol. **65**(3), 1098–1108 (2015)
13. Zhang, L., Jiang, S., Huang, K., et al.: Knowledge graph-based network analysis on the elements of autonomous transportation system. In: 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 536–542. IEEE (2021)
14. Deng, Z., Xiong, C., Cai, M.: Research on theoretical model and construction method of the physical object for autonomous transportation system. In: CICTP 2022, pp. 647–655 (2022)
15. Deng, Z., Cai, M., Xiong, C.: An architecture integrity simulation evaluation method for an autonomous transportation system based on an information-triggered collaboration mechanism. IEEE Intell. Trans. Syst. Mag. https://doi.org/10.1109/MITS.2023.3272501
16. Younis, O., Moayeri, N.: Employing cyber-physical systems: dynamic traffic light control at road intersections. IEEE Internet Things J. **4**(6), 2286–2296 (2017)
17. Guzmán, J.A., Núñez, F.: A cyber-physical systems approach to collaborative intersection management and control. IEEE Access **9**, 99617–99632 (2021)
18. Kamal, M.A.S., Tan, C.P., Hayakawa, T., et al.: Control of vehicular traffic at an intersection using a cyber-physical multiagent framework. IEEE Trans. Ind. Inf. **17**(9), 6230–6240 (2021)

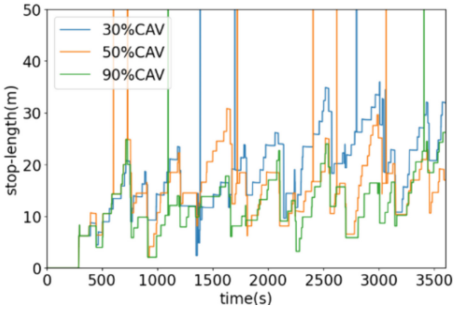


Fig. 6. Comparison of queue length under Webster adaptive control under different CAV penetration rates

of research on the application and modeling problems of CPS in the transportation industry, and identifies the key problems in the development of TCPS: the lack of complete and effective theoretical guidance and the lack of scenario-level TCPS modeling.

In order to solve this problem, this paper, with the help of the theoretical framework of ATS, starts from the five categories of ATS: service, technology, demand, function and technology, and focuses on the connection of function, transmission of data flow, and connection of entity to form the three categories of functional, logical and physical architectures respectively, and analyzes the typical traffic scenarios according to this idea to form a complete scenario architecture.

Subsequently, under the guidance of intersection scenario architecture, this paper draws on the design of cyber layer and physical layer of information-physical system, and adopts Netlogo, a multi-intelligence approach software, to simulate physical objects, information flow and information interaction pairs in the scenario architecture respectively, and constructs the physical layer of intersection scenario; adopts Python language for joint simulation, and designs the flow change process of information flow in the cyber layer. After the model was successfully constructed, it was simulated and the simulation results were analyzed qualitatively for functional integrity and quantitatively for traffic indicators.

However, the current study still has certain shortcomings. First, in the physical layer of the simulation model, simulation software as well as variables are still used to simulate the real world in the physical layer, which cannot really simulate the variability and complexity of the real world. Second, in the cyber layer of the information out process, there are still areas that can be improved. The next step is to combine the work of this paper with real-world intersection data, compare the differences between this model and the actual data, and use it to continuously adjust the model to form a more accurate and more reflective real-world intersection TCPS model.

References

1. Chen, H., Cai, M., Huang, K., et al.: Classification and evolution analysis of key transportation technologies based on bibliometrics. Sci. Program. **2021**, 1–13 (2021)

Table 3. Traffic participants and usage information flows in similar literature

Author of the literature	Traffic participants in the model	The flow of information that the model leverages
Younis O	Cars and Signal Lights	Vehicle coordinates and speed information, traffic flow information
Guzman J A	Cars and Signal Lights	Vehicle coordinates and speed information, traffic flow information
Kamal M A S	Cars and Signal Lights	Vehicle coordinates and speed information, traffic flow information
This article	Cars and Signal Lights, Roadside infrastructures, traffic operation center, meteorological information center	Vehicle coordinates and speed information, traffic flow information, road surface information, weather condition information, road congestion information

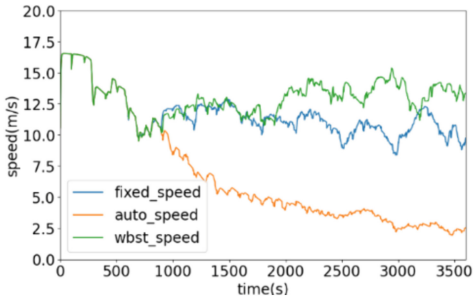


Fig. 5. Comparison of average vehicle speed of three types of phase control methods with 30% CAV

It can be found that CAV has a significant improvement in queue length at higher penetration rates, but the improvement is not significant enough, which may be related to the fact that the control method adopted lacks vehicle-road cooperative driving and cannot fully utilize the potential of CAV traffic.

5 Conclusion

This paper reviews the key technologies that have helped the rapid development of the transportation industry in recent years, and provides an overview of CPS, which has gradually emerged with the development of various technologies, especially the current state

4.4 Simulation Conclusion

1. Integrity Analysis

During the simulation, the real-time operation of CPS can be simulated, and the integrity of the operation of information flow such as weather information, vehicle information, and infrastructure information can be properly demonstrated respectively. The integrity test of the control schemes such as vehicle speed limit and no left turn can operate normally. The following figure shows some of the information flows involved in the intersection scenario and the results of real-time information access during the simulation (Table 2).

Table 2. Simulation test results of perceptual information flow in scenario architecture

Information Flow	Form of embodiment	Test results
Environment information	Every 60 s, the Weather Center perceives environmental information	Normal
Traffic flow basics	Average speed, queue length	Normal
Traffic flow information	Traffic flow	Normal
Lane monitoring data	Zoning according to roadside facilities (Location and time of infrastructure failures)	Normal
Road guardrail monitoring data		
Road sign marking monitoring information		
Vehicle driving condition monitoring information	Whether the vehicle is faulty (Fault number and time)	Normal
Vehicle location and movement information	Vehicle location information	Normal

The following table compares the intersection TCPS functions designed in this paper with similar literature [16–18] (Table 3):

2. Traffic Indicator Analysis

The model proposed in this paper can also be applied to traditional traffic microsimulation, and the diversity of agents in it can help support traffic flow simulation in future heterogeneous mixing phases.

The following results were obtained by simulating different signal control schemes for CAV penetration of 30% at a traffic volume of 360veh/h and analyzing the average vehicle speed when different signal control schemes were used (Fig. 5):

It is easy to see that the traffic Indicator based on Webster's adaptive control method performs better.

Then compare the queue length using the same Webster's adaptive control method for different CAV penetration rates, as shown below (Fig. 6):

Different types of agents possess different variables and model the differences in information flow with changes in variables. The following table briefly describes some of the variables belonging to each agent and the meaning of these variables (Table 1).

Table 1. Agents and some of their variables

Agent type	Variable	Meaning
Cars	Xcor, ycor	The location of the vehicle
	v	Vehicle speed
	a	Vehicle acceleration
	Auto-type	Decide whether the vehicle is CAV or HDV
	Driver-type	Decide on the type of HDV driver
	Normal	Determine the condition of the vehicle's interior
Weather Center	Weather	Weather conditions collected every minute
Roadside facilities	Congestion	Congestion on roads near roadside facilities
Signal Lights	Sign	Determines the phase of the signal lights
Transportation Operations Center	Control	Decide on the control measures to be taken by the transportation operations center

4.3 Simulation Experiments

After defining the cyber layer and physical layer of the scenario separately, the corresponding simulation experiments can be taken to analyze the practicality and effectiveness of the model.

The simulation experiments focus on two parts: the examination of the scenario model on the integrity of information flow and physical objects described by the ATS architecture, and the comparative analysis of various traffic flow metrics, including average speed and queue length for various CAV penetrations under different phase control methods. The simulation time step is taken as 0.1 s, and the simulation length is 3600 time steps.

response" according to the different ways of processing and utilizing information. These four steps realize the initial processing of data through built-in data calculation formulas; the generation and comparison of advantages and disadvantages of decision solutions through built-in decision algorithms; and the effective use of data and effective feedback to the physical world through built-in control solutions. In the actual simulation process, the pynetlogo module built in Python is used to co-simulate with the Netlogo used for physical layer simulation, and the communication, computation and control functions of data are simulated by different operations respectively (Fig. 4).

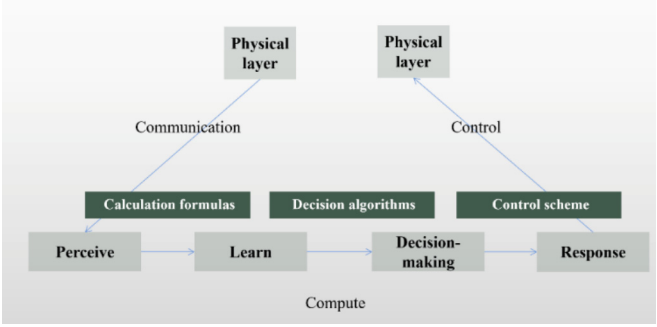


Fig. 4. Schematic diagram of the cyber layer structure

4.2 Scenario Overview

In the actual simulation experiment, the intersection scenario applied to the common four inlet lane intersection, where each inlet lane contains three lanes of left turn, straight ahead, and right turn, and the scenario contains five major categories of traffic participants: vehicles, signal lights, roadside infrastructures, weather center, and traffic operation center.

Vehicles are divided into two types of Connected-Automated Vehicle (CAV) and human driving vehicles (HDV), which take three different categories of follow-the-road models, IDM, CAC and CACC, according to their own vehicle types, where HDVs are divided into aggressive, normal and conservative types according to the characteristics of the drivers, and take different reflection times during the driving process according to the driver types. Vehicles will also monitor their own conditions in real time and take countermeasures in case of abnormal conditions, with vehicle arrivals conforming to Poisson distribution and a flow rate of about 360 vehicles per hour; Signal control methods include four-phase fixed-cycle, adaptive phase control based on the traffic volume of the previous cycle, and adaptive phase control based on Webster's timing method; The Roadside infrastructures will sense road congestion and road conditions in real time, and use them to impose speed limit notices and other measures on vehicles; the weather center will collect weather information at regular intervals and distribute it to vehicles and drivers; the traffic operation center can take common intersection control measures such as speed limit and no left turn according to the actual situation.

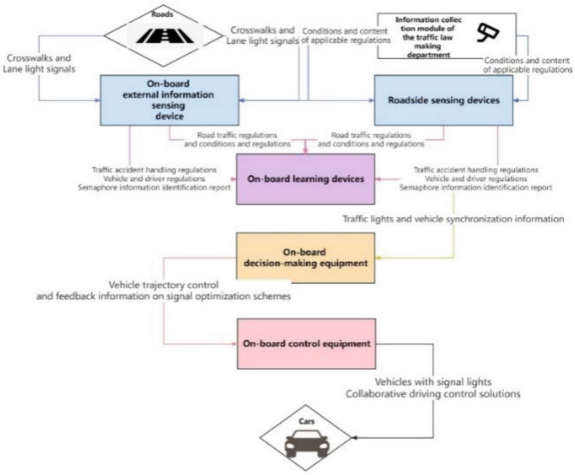


Fig. 3. Physical architecture of vehicle and signal light coordination sub-service in the intersection scenario

4 Modeling Based on ATS Scenario Theory

After the introduction of the theoretical framework of ATS, it is necessary to consider how to carry out scenario-oriented TCPS construction under the guidance of relevant theories, and to simulate and test the model after its successful construction in order to judge whether the model has good completeness and practicality.

4.1 Model Design

During the construction of the model, the structural relationships within the physical layer and the cyber layer need to be designed separately. First is the physical layer, which helps to examine the completeness and richness of the scene model with the help of the scenario architecture theory of ATS. The scenario architecture contains several kinds of physical objects that interact and influence each other, and the multi-agent simulation software Netlogo is used for the construction of the physical layer for this feature. In the mapping process of the simulation software to the scenario architecture, the basic elements in these three types of scenario architectures are mapped into the simulation software as the physical objects, the variables and global variables of the agents as the information flow, and the interaction between the agents and the data transmission process as the information interaction process, respectively, to realize the design of the physical layer.

The cyber layer is mainly concerned with the whole process of information from generation to being perceived, processed and utilized, so as to realize the effective application of real-time data in the cyber physical system, and therefore the cyber layer is designed as shown in the figure below. Information is generated in the physical world and reaches the cyber layer through V2X and other communication technologies. The cyber layer is divided into four steps: “perception → learning → decision making →

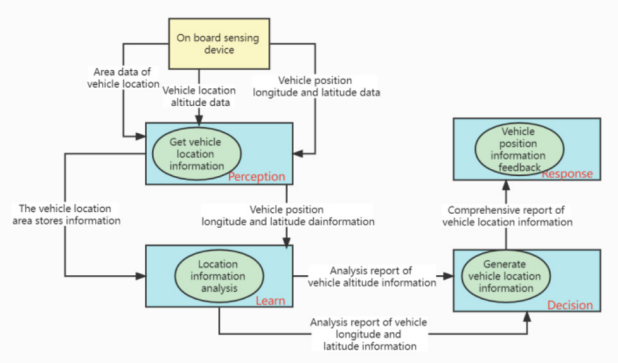


Fig. 2. Logical architecture of vehicle location aware service

they are clustered into information streams, and the information streams are connected in the physical objects with the help of the basic properties of the physical objects, and finally the physical architecture is generated as shown in the figure below.

3.3 Architectures of the Scenarios

ATS has built and analyzed the physical architecture for five typical scenarios, specifically divided into MaaS scenario, Electric Bus Operation scenario, Cargo Multimodal transportation scenario, Highway Formation scenario, and Intersection self-driving vehicle pedestrian avoidance scenario. Similar to the physical architecture of sub-services, the scenario architecture is also composed of three types of elements: physical objects, information flow and information interaction pairs.

In terms of concrete implementation, the demands and components contained within the scenario are analyzed by collecting a large number of definitions related to the scenario, and this is used as a guide to gradually obtain the services and functions that the scenario needs to provide, as well as the technologies needed to achieve them; subsequently, the corresponding three types of architectures correspond to the required sub-services, and on the basis of these materials, the physical architecture of the scenario is constructed with the help of a collaborative mechanism of elements and architectures [14, 15]. As an example, the intersection self-driving vehicle pedestrian avoidance scenario contains 27 sub-services and 117 sub-functions (Fig. 3).

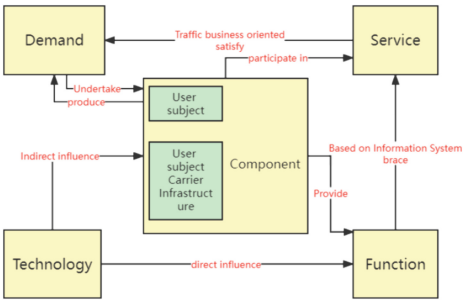


Fig. 1. Contact diagram of five types of ATS elements

3.2 Three Types of Architectures

To better explain the correlation of functions in each ATS service and to guide the construction of realistic infrastructure, it is necessary to form corresponding functional, logical and physical architectures for each service.

The ATS functional architecture serves as the initial architecture to describe the linkage between functions and to determine the logical sequence for implementing the functions in order to guide the subsequent architecture. Specifically, through the implementation logic of "sense, learn, decide, respond", a number of functions involved in each sub-service are classified, and the functions are divided into sub-functions according to the implementation process, and finally the sub-functions are connected in the order of their implementation in the service, and finally the functional architecture is formed.

ATS logical architecture is based on the understanding of the traffic semantics of the service, which is realized by organizing the information system functions, and plays the role of connecting "traffic" and "information functions", which is manifested in two aspects. On the one hand, it can express the hierarchical and progressive relationship between functions, that is, the further expression of "perception-learning-decision-response", and form a hierarchical system between functions, which provides a theoretical basis for architectural reconstruction, integration and optimization in any scenario. For this purpose, we layered perception into acquisition and identification, learning into fusion and analysis, decision making into generation of solutions and selection of optimal solutions, and corresponding into execution and feedback. On the other hand, the logical architecture must also clarify the input-output relationship of each function, and thus express the process of service implementation (Fig. 2).

The physical architecture is the carrier for the transformation of the logical architecture to the real transportation entity, the framework view that guides the planning and construction of the transportation system, and the ultimate embodiment of the basic theory of ATS. Firstly, the system functions in ATS need to be mapped to real traffic entities, and for this purpose, according to the current traffic system construction, the theoretical model of "physical object" is proposed, and its properties are described in terms of ontological attributes and connectivity attributes, the former including object categories, autonomy, and traffic information participation methods, and the latter including access capability and mapping logic [14]. Subsequently, by analyzing the types of data streams,

the transportation industry, which refers to the construction idea of CPS to establish the mechanism of sensing, communication, computation, and application of traffic information, so as to achieve the improvement of traffic efficiency and the efficient control of vehicles and infrastructure [9]. The current research on TCPS is divided into several aspects, some scholars focus on the computation and communication time problems of TCPS, and compare the advantages of TCPS through the time and efficiency of information transmission [10]; some scholars focus on the data security and related problems of TCPS, and propose various methods to guarantee the data security of the system [11]; some scholars focus on the modeling problems of TCPS, abstractly model the traffic events existing in the real world, and analyze the advantages and problems of TCPS through simulation.

The modeling of TCPS includes traffic modeling at the physical layer and network modeling at the cyber layer [12]. Due to the lack of real-world actual data of TCPS, the modeling of the physical layer of TCPS often relies on common microscopic traffic simulation software such as Sumo and Vissim. However, TCPS, as a complex fusion system, is concerned with the impact of a wide variety of information flows on the actual traffic, but most current studies rely on microscopic traffic simulation software, which can only simulate common traffic participants such as vehicles and signals, and cannot achieve effective simulation of the physical layer. Research on modeling the cyber layer for TCPS can be divided into two categories, one that models the cyber layer by considering vehicles as communication nodes and analyzing the communication metrics of multi-node networks, and the other that focuses on different decision and control algorithms to build the cyber layer of CPTS by modeling out the generation, processing, and analysis of data.

In general, there are still few studies on modeling TCPS as a trend of traffic system development, and there are drawbacks such as low completeness and lack of theoretical guidance, thus this paper decides to model and analyze the intersection scenario of TCPS with the help of the complete theoretical framework of ATS.

3 ATS Theoretical Framework

3.1 Five Types of Elements

ATS has been designed using object-oriented design methodology with 5 categories of basic elements of ATS. The capabilities possessed by the transportation system are decoupled into a number of mutually independent basic units, which are services, the transportation tests and requests made to the transportation system are summarized as requirements, the factors that facilitate the capabilities and evolution of the transportation system are summarized as technologies, the units that realize the capabilities of the transportation system are summarized as functions, and the participating roles of the transportation system are summarized as components, which are related as shown in the following figure [13] (Fig. 1).

for the common intersection scenarios in transportation system, the physical world and cyber space under this scenario are designed respectively, and the intersection scenario model based on ATS theory is constructed through the definition of information transmission and application methods, and simulation analysis is conducted for the completeness of the scenario and each traffic index of the scenario to form a complete TCPS scenario modeling technology system, which is conducive to guiding the development of future traffic information physical systems.

2 Related Works

2.1 CPS and Modeling

CPS enables the perception and control of the physical world through the integration of the physical world and the in cyber space, using advanced perception, communication, and computing technologies, with strong real-time capabilities. In recent years, countries around the world have been investing a lot of efforts in CPS research, and in 2022, the Chinese Natural Science Foundation listed CPS as one of 115 "priority areas for development".

The key step of CPS from abstract architecture to concrete model is the modeling of CPS, and the problem is the focus and difficulty of CPS research and has attracted the attention of a large number of scholars worldwide. The modeling process is limited by the characteristics of discrete CPS cyber layer, continuous physical layer, containing more elements, and the integration of multiple industrial fields, which cannot use the traditional modeling method and needs to take into account multiple aspects. The current solutions to the problem fall into several categories: First, the traditional discrete system modeling methods are borrowed to build discrete systems, mainly including formal modeling and high-level language modeling. Among them, formal modeling includes formal inference modeling, extended Petri net modeling, time automaton modeling and other methods, while high-level language modeling includes AADL, modelica, UML, etc., and reaches the unification of the overall system through the discretization of continuous events; Second, drawing on the traditional continuous system, model the continuous system in CPS with the help of traditional continuous system modeling methods, such as parametric model, Newtonian mechanics, etc., and then reasonably embed the discrete events into the continuous system to achieve the unification of the two; The third approach is hybrid modeling, where CPS modeling of hybrid tools is achieved by modeling the physical and cyber layers separately and disposing the interfaces between them rationally [7]. In addition, some scholars try to adopt emerging technologies such as group intelligence [8] and data-driven to solve the problem, but they are limited by the maturity of technology development and need further research and improvement.

2.2 TCPS and Modeling

In recent years, with the increasing traffic demand and the deep application of traffic data, a physical system of traffic information combining 3C (communication, computation, and control) technology and traffic elements has become a hot research topic in

technology infrastructures, and the improvement of service quality due to technological developments, but also in stimulating the emergence of more demanding transportation demands [1].

Specifically, the advent of next-generation data communication technologies makes larger data transfers and very small communication delays possible, which allows transportation systems to obtain the state of traffic participants (people, vehicles, roads and environment) in near real-time or quasi-real-time conditions, greatly enhancing the sensing capabilities of transportation systems. Meanwhile, with the widely application of AI, ML, and IC technologies, the computing capacity of end-users and edge devices is fully utilized, making it possible for end-users to obtain their own status with high precision in most cases, which not only reduces the amount of information transmission, but also supports more personalized and intelligent system services. In addition, the development of distributed system management technology, optimization theory and other technologies make the collaboration process among the agents more robust and efficient, and also play a strong guarantee in terms of system scalability and security. The development of these three types of technologies has led to the efficient implementation of the three main functions of sensory and communication, computing and processing, decision and control of data in transportation system, respectively, which are the main components of concern for transportation cyber physical system (TCPS) [2]. The transportation cyber physical system regards the real traffic world as the physical entity layer, and transmits all kinds of traffic data generated by the physical entity layer, such as traffic flow, vehicle speed, parking delay to the cyber space in real time, and generates some kinds of control schemes in real time through many kinds of data processing and decision algorithms, and synchronizes them to the physical world to implement control, so as to realize the effective utilization of traffic information.

At the same time, emerging technologies are transforming transportation systems from "passive" to "active" in meeting the demands of transporting people and goods, which has given birth to the concept of autonomous transportation systems (ATS). ATS realizes transportation by self-organized operation and autonomous service. Its operation logic is autonomous perception, autonomous learning, autonomous decision making and autonomous response, which is essentially to reduce human intervention in transportation system and enhance the autonomous capability of transportation system, specifically in the four aspects of active response to traffic demand, automatic operation of vehicles, active control of infrastructure and active adaptation of external environment. With the help of system engineering modeling theory, a set of theoretical framework of transportation system from transportation elements to specific guiding structures has been constructed with "autonomy" as the core in the study of ATS [3–5].

Among the studies targeting TCPS, simulation techniques try to reproduce the implementation logic of TCPS and are one of the main techniques related to TCPS. However, due to the synergistic requirements of TCPS for multiple domains and the lack of effective theoretical guidance, the current research on TCPS simulation technology is more focused on the study of modeling methods and mainly stays at the level of individual events [6].

To this end, this paper first analyzes the development status of TCPS, and then introduces the theoretical framework of ATS from elements to architecture to scenarios. Next,

Cyber Physical System Modeling and Analysis in Typical Scenarios Based on the Theory of Autonomous Transportation System

Zi-Sheng Zhou[1,2], Ming Cai[1,2], Zhuo-Lin Deng[1,2], and Chen Xiong[1,2(✉)]

[1] School of Intelligent Systems Engineering, Sun Yat-Sen University, Guangzhou 510006, China
xiongch8@mail.sysu.edu.cn

[2] Guangdong Key Laboratory of Intelligent Transportation Systems (ITS), Guangzhou 510006, China

Abstract. With the rapid development and application of sensing, computing and controlling technologies in the transportation industry, the construction of transportation cyber physical system (TCPS) relying on these three types of technologies has gradually become a research hotspot in the transportation system. However, the modeling of TCPS suffers from a lack of theoretical guidance and a single modeling hierarchy. To this end, this paper introduces the theoretical framework system of Autonomous Transport System (ATS) as the theoretical basis for TCPS scenario modeling, and sorts out the theoretical framework of ATS from five types of elements to three types of architectures to typical scenarios. Then, taking the modeling of TCPS in intersection scenario as an example, the physical layer of the model is constructed by mapping the physical objects, information flow, and information interaction pairs in the ATS scenario architecture, and the cyber layer of the model is designed by the service implementation logic of ATS, and through the process of data generation and application to achieve the integration of the two layers of applications. After the model was constructed and simulations were implemented, the functional integrity of the scenario reflected by the model was analyzed qualitatively, and the results of specific traffic indicators under different parameters were analyzed quantitatively to verify the integrity and validity of the model.

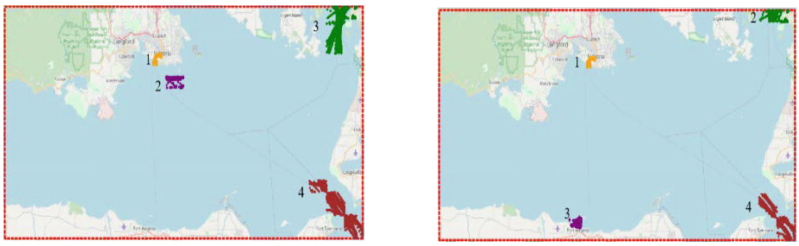
Keywords: Cyber-physical system · traffic system modeling · autonomous transportation system

1 Introduction

With the rapid development of information and intelligent technology, the new generation of communication technology, Internet of Things, big data, AI, mobile Internet, energy management and Intelligent & connected vehicle technology are gradually used in the field of intelligent transportation system. This is manifested not only in the rapid growth of passenger volumes, the orderly construction and replacement of information

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 165–177, 2024.
https://doi.org/10.1007/978-3-031-65126-7_17

2. Ning, Z., et al.: Blockchain-enabled intelligent transportation systems: a distributed crowd-sensing framework. IEEE T Mob. Comput. **21**(12), 4201–4217 (2021)
3. Ning, Z., et al.: Intelligent resource allocation in mobile blockchain for privacy and security transactions: a deep reinforcement learning based approach. Sci. China Inf. Sci. **64**(6), 162303 (2021)
4. Ning, Z., et al.: Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE T Mob. Comput. **22**(5) (2021)
5. Ning, Z., Chen, H., Ngai, E.C., Wang, X., Guo, L., Liu, J.: Lightweight imitation learning for real-time cooperative service migration. IEEE T Mob. Comput. (2023)
6. Ferreira, M.D., Campbell, J., Purney, E., Soares, A., Matwin, S.: Assessing compression algorithms to improve the efficiency of clustering analysis on AIS vessel trajectories. Int. J. Geogr. Inf. Sci. **37**(3), 660–683 (2023)
7. Egala, B.S., Pradhan, A.K., Badarla, V., Mohanty, S.P.: Fortified-chain: a blockchain-based framework for security and privacy-assured internet of medical things with effective access control. IEEE Internet Things **8**(14), 11717–11731 (2021)
8. Xiong, L., Xiong, X., Zhang, F., Chen, H.: Unsupervised Deep Embedding Clustering for AIS Trajectory. In: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 2022, pp. 2283–2286 (2022)
9. Deebak, B.D., Al-Turjman, F.: Smart mutual authentication protocol for cloud based medical healthcare systems using internet of medical things. IEEE J. Sel. Area Commun. **39**(2), 346–360 (2020)
10. Ali, A., et al.: Security, privacy, and reliability in digital healthcare systems using blockchain. Electronics **10**(16), 2034 (2021)
11. Bureva, V., Popov, S., Traneva, V., Tranev, S.: Generalized net model of cluster analysis using CLIQUE: clustering in quest. Int. J. Bioautom. **23**(2), 131 (2019)
12. Wang, S., Li, Y., Xing, H.: A novel method for ship trajectory prediction in complex scenarios based on spatio-temporal features extraction of AIS data. Ocean Eng. **281**, 114846 (2023)
13. Ganesh, E.N., Rajendran, V., Ravikumar, D., Kumar, P.S., Revathy, G., Harivardhan, P.: Detection and route estimation of ship vessels using linear filtering and ARMA model from AIS data. Int. J. Oceans Oceanogr. **15**(1), 1–10 (2021)
14. Hartawan, I.P.N., Widyantara, I.M.O., Karyawati, A.A.I.N.E., Er, N.I., Artana, K.B., Sastra, N.P.: AIS data pre-processing for trajectory clustering data preparation. In: 2021 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), Bali, Indonesia, 2021, pp. 1–5. (2021)



(a) Areas with frequent ship movements on January 1

(b) Areas with frequent ship movements on January 2

Fig. 4. Visualization of the map of the frequent movement area of ships with a single spatiotemporal granularity

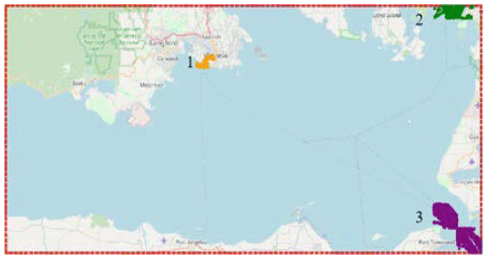


Fig. 5. Visualization of the map of the frequent movement area of ships with a multi-temporal and spatial granularity based on two-day data from January 1st to 2nd

spatial-temporal granularity shows great differences on different dates, which proves the effectiveness of the method proposed in this paper.

4 Conclusion

In this paper, we propose a method for extracting frequent ship moving areas based on grid density peak clustering, which solves the problem that grid density peak clustering methods need to manually select cluster centers. To learn more fine-grained spatial-temporal information, we consider frequently active regions of both spatial and temporal information. We fuse the frequently active regions with single spatiotemporal granularity on the timeline to obtain frequent active regions with multiple spatiotemporal granularities, which makes the extracted frequent active regions more accurate. In simulation experiments, we evaluate the effectiveness of the proposed ship frequent activity area extraction method and compare it experimentally with other methods. The results show that our method can more accurately and effectively extract the areas with frequent ship activities.

References

1. Wang, X., Ning, Z., Guo, L., Guo, S., Gao, X., Wang, G.: Mean-field learning for edge computing in mobile blockchain networks. IEEE T Mob. Comput. **22**(10), 5978–5994 (2022)

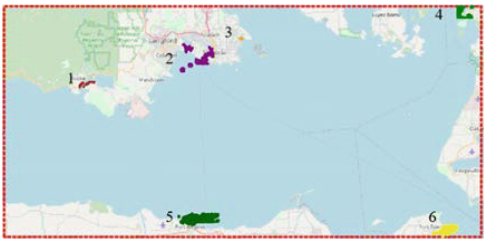


Fig. 2. Map visualization of frequent activity areas by using CLIQUE method

area 2, but it is identified as a frequent activity area. However, some areas among the screened-out areas have very high ship densities, such as grid 18, grid 38, grid 399, grid 379 and grid 358, which represent areas with significantly higher ship density than other areas., but was identified as an infrequently active area.} Fig. 3 shows the frequently active regions extracted by grid density peak clustering. Compared with Fig. 2, the density of frequent active areas extracted in Fig. 3 is in the forefront, which shows the effectiveness of the advanced selection threshold method proposed in this paper, which can correctly screen out high-density grids.

Fig. 3. Map visualization of frequent activity areas by using AGDPC method

Figure 2 and Fig. 3 only consider the spatial information of the frequent activity area extraction method, and can only obtain the frequent activity area in the entire large time period, but cannot obtain the frequent activity area in different time periods. In order to extract frequent activity areas more accurately, this paper firstly extracts the frequent activity areas of ships with single spatio-temporal granularity in different time periods in the same area under the given time granularity and spatial granularity. On this basis, on the time axis, if there is an intersection between frequent ship activity areas in adjacent time periods, the frequent activity areas in adjacent time periods are merged. Otherwise, the fusion of the next time period is performed until the time span is traversed.

The frequently active regions extracted at a single spatio-temporal granularity using the grid density peak clustering method are shown in Fig. 4. And the frequent movement area of ships with multiple spatial and temporal granularities is shown in Fig. 5. It can be seen that the frequent ship activity areas under multiple spatial-temporal granularity tend to be consistent on different dates, and the propagation activity area of a single

3.1 Experimental Parameter Settings

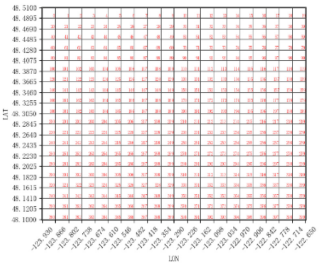
The parameter setting of the comparison method in the experiment is selected through manual tuning, as shown in Table 1.

Table 1. .

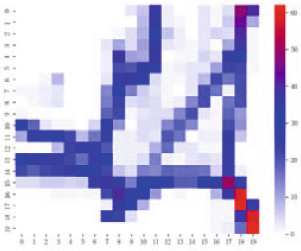
Parameter	Value
Meshing	20*20
Density threshold	200
Time granularity	1(day)

3.2 Results and Analysis

We first divide the experimental data into 20*20 grid areas. The number of ships in each grid area and the heat map are shown in Fig. 1. Subsequent experiments will be compared with Fig. 1.



(a)Meshing and grid number



(b) heat map of ship distribution in grid

Fig. 1. The attributes of the research area (a) Meshing and grid number, (b) heat map of ship distribution

The extraction of frequent ship activity areas at a single spatio-temporal granularity refers to extracting frequent ship activity areas in different time periods within the same research area, given the time granularity and spatial granularity. We divide the time range into several uniform equal parts, and divide the space range into m*m grids. An improved grid density peak clustering algorithm is used to automatically select the cluster center and extract its frequent activity areas at a single spatio-temporal granularity. Figure 2 shows the frequent activity areas of ships extracted by CLIQUE. It extracts 6 frequently active regions. However, compared with Fig. 1, it can be seen that the ship density in some of the six frequent activity areas is very low, such as grid 107 in area 1, area 3, and

identified. First, count the number of times that each mesh object in the cluster is the nearest higher-density mesh object to other mesh objects:

$$nt_j = \sum_{i=1}^{n} z\left(j - \underset{j:\rho_j > \rho_i}{arg\,min}\left(d_{ij}\right)\right) \tag{1}$$

In the formula, $z(x) = \begin{cases} 1, x = 0 \\ 0, other \end{cases}$. d_{ij} is expressed as the Euclidean distance between the grid object i and the grid object j. ρ is the local density of the mesh object. Since it is difficult for a point located on the boundary of a cluster to become the closest high-density mesh object to other mesh objects, when is 0, the mesh object is usually located on the boundary area, so the core area of the cluster can be defined as:

$$c_{core}^{k} = \left\{x_i | \rho_i > \max\left(\rho_j\right), x_i \in c^k, x_j \in c^k \& nt_j = 0\right\} \tag{2}$$

In the formula, c^k represents the clusters obtained by clustering, c_{core}^{k} represents the core region of the class cluster c^k, $\max(\rho_j)$ is the maximum function. In these cluster core areas, although their density is larger than their neighbors, from the overall data distribution, some areas have relatively few ships and should not be considered frequent activity areas. In order to obtain the frequent activity areas of ships that meet the actual situation, these core areas need to be further screened. In order to reduce human participation, this paper automatically selects the density threshold $d_{th} = \max_{ntj=0}(\rho_j)$ according to the distribution characteristics of grid density in various clusters, and selects the grids whose grid density exceeds the threshold in various clusters:

$$area_{fre} = \left\{x_i | \rho_i > d_{th}, x_i \in c_{core}^{k}\right\} \tag{3}$$

By merging adjacent high-density mesh objects, the ship frequent activity area can be obtained. However, the ship frequent activity area extracted by this method ignores the time information. In fact, the areas of ship activities are different at different times. In this paper, by fusing ship frequent areas with single spatio-temporal granularity on the time axis, more accurate ship frequent areas with multiple spatio-temporal granularities are obtained.

3 Experiment and Analysis

In order to validate the proposed method, this paper uses two common frequent activity region detection methods for comparison. The first is the classic grid clustering method Clustering In QUEst(CLIQUE) [9], which uses the number of data points in the grid as the grid density to extract areas with frequent ship activities; the second is the advanced grid density peak clustering method proposed in this paper. This experiment selected AIS data of ships in the sea area of 122°35′W−123°55′W, 48°06′W−48°30′N from January 1, 2019 to January 3, 2019, and the data comes from the open source website https://marinecadastre.gov/ais/.

such as Maritime Mobile Service Identity (MMSI), vessel position, and speed. By analyzing and mining AIS data, extracting frequent activity areas of vessels can provide technical support for research in detecting abnormal vessel behavior, predicting port traffic flow, voyage planning [11], and recognizing maritime target intentions.

At present, clustering algorithms have been widely used in the research of object hotspots region extraction [12]. Wang et al. [13] developed a rapid clustering model of trajectories based on hierarchical modeling. Each ship state establishes its trajectory similarity model and performs recursive clustering of ship trajectories from top to bottom, avoiding the cumbersome calculation of existing ship clustering models, high time complexity, difficult parameter adjustment process, and other shortcomings. In [14], a spectral clustering algorithm is used to cluster the sub-trajectory segments to identify representative ship maneuvering behavior trajectories. Hartawan et al. [11] suggested that a typical motion model of ships in the area could be obtained based on AIS data by DBSCAN clustering of ship trajectory segments combined with track similarity measure and extraction of typical trajectories.

The above methods only focus on the spatial information of moving objects and ignore the time information, which will result in the identification of an area with no ships or only a few ships in a certain interval as an area where ships are frequently active. In this regard, this paper proposes an advanced grid density peak clustering method (AGDPC). This method can extract frequent ship motion areas at multiple time granularities while using spatial clustering and considering ship time information.

2 Advanced Grid Density Peak Clustering

Traditional grid density peak clustering requires manual determination of cluster centers, which can easily lead to inaccurate clustering results. To address this problem, this paper uses the boxplot method and the elbow method to automatically determine the cluster center and number of clusters. First, in order to solve the problem that the local density and relative distance of grid objects affect the selection of cluster centers due to different dimensions, this paper first uses the minimum and maximum normalization method to map the value range of grid objects to the local density and relative distance of grid objects. Perform normalization processing between 0 and 1, and preselect the cluster center set. Then, use the box plot method to calculate its upper and lower bounds and quartiles to obtain its box plot distribution. Grid objects with higher local density and relatively far distance is further filtered according to the box plot distribution as a cluster set. At this time, the cluster center candidate set may contain more cluster centers than the actual situation, causing the classification of clusters to be too detailed. Because there may be grid objects in the cluster set that have a high local density but a small relative distance, or a grid object that has a small local density but a large relative distance, it is necessary to further screen the cluster center candidate set. This paper uses the elbow method to filter the cluster set. By finding the inflection point of the cluster center, the candidate points before the inflection point are used as the cluster center, and the remaining candidate points are assigned to the same cluster as its nearest high-density neighboring grid object., complete the cluster analysis of ship AIS data and obtain the ship activity area.} Based on the above ideas, the core regions of the clusters can be

Extraction of Frequently Active Areas of Ships Based on Advanced Grid Density Peak Clustering

Xuanrui Xiong[1], Han Shen[1], Lanke Zhu[2,3], and Jianbo Zheng[2,3](✉)

[1] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
xiongxr@cqupt.edu.cn, s210131206@stu.cqupt.edu.cn

[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{1120200314,jianbo.zheng}@smbu.edu.cn

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

Abstract. The cognition of the frequent activity areas of ships based on AIS data is of great significance in reducing port navigation risks and improving the efficiency of ships entering and leaving ports. Traditional extraction methods only consider spatial information and ignore the impact of temporal information on clustering results, resulting in inaccurate extraction of frequently active areas. We propose an advanced grid density peak clustering method (AGDPC) to extract frequently active areas, which can advanced select cluster centers and density thresholds to solve the problem that grid density peak clustering methods cannot advanced select cluster centers. The improved grid density peak clustering method is used to extract frequent ship motion regions under a single spatial-temporal granularity according to a given spatial-temporal granularity. Then, we fuse multiple ship frequent activity areas to obtain multi-temporal and spatial granularity ship frequent activity areas. Experimental results show that this method can extract frequent motion are-as more accurately than traditional methods, and better reflect the ship's navigation rules.

Keywords: Trajectory clustering · Grid density peak clustering · Frequent activity areas extraction · AIS data

1 Introduction

The mining and analysis of existing data is one of the important means to predict and evaluate the future situation of objects. With the development of machine learning and deep learning, data mining and analysis techniques are widely used in economics, edge computing [1–3], blockchain [4–7] and other fields [8–10]. The Automatic Identification System (AIS) is a type of ship navigational system that contains essential information

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 158–164, 2024.
https://doi.org/10.1007/978-3-031-65126-7_16

7. Faragallah, O.S., et al.: Efficiently encrypting color images with few details based on RC6 and different operation modes for cybersecurity applications. IEEE Access **8**, 103200–103218 (2020)
8. Farooq, M.U., Waseem, M., Mazhar, S., Khairi, A., Kamal, T.: A review on internet of things (IoT). Int. J. Comput. Appl. **113**(1), 1–7 (2015)
9. Ghadirli, H.M., Nodehi, A., Enayatifar, R.: An overview of encryption algorithms in color images. Signal Process. **164**, 163–185 (2019)
10. Gokhale, P., Bhat, O., Bhat, S.: Introduction to IoT. Int. Adv. Res. J. Sci. Eng. Technol. **5**(1), 41–44 (2018)
11. Grichi, M., Abidi, M., Jaafar, F., Eghan, E.E., Adams, B.: On the impact of interlanguage dependencies in multilanguage systems empirical case study on java native interface applications (JNI). IEEE Trans. Reliab. **70**(1), 428–440 (2020)
12. Hwang, S., Lee, S., Kim, J., Ryu, S.: Justgen: effective test generation for unspecified JNI behaviors on JVMs. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pp. 1708–1718. IEEE (2021)
13. Jihui, Y., Qinian, Z., Zhenhao, Z.: Cloud storage and security technology based on the the internet of things. ZTE Technol. J. **18**(6), 12–16 (2012)
14. Lee, S., Lee, H., Ryu, S.: Broadening horizons of multilingual static analysis: semantic summary extraction from c code for JNI program analysis. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, pp. 127–137 (2020)
15. Li, B., Feng, Y., Xiong, Z., Yang, W., Liu, G.: Research on AI security enhanced encryption algorithm of autonomous IoT systems. Inf. Sci. **575**, 379–398 (2021)
16. Liu, F., Koenig, H.: A survey of video encryption algorithms. Comput. Secur. **29**(1), 3–15 (2010)
17. Lu, H., Jin, C., Helu, X., Zhu, C., Guizani, N., Tian, Z.: AutoD: intelligent blockchain application unpacking based on JNI layer deception call. IEEE Netw. **35**(2), 215–221 (2020)
18. Manikandan, V., Amirtharajan, R.: On dual encryption with RC6 and combined logistic tent map for grayscale and DICOM. Multimed. Tools Appl. **80**(15), 23511–23540 (2021)
19. Ming, H., Jun, J., Xiaohu, C., Guohua, C.: Technology and security of internet of things. Comput. Secur. **4**, 49–52 (2011)
20. Raghavan, B., Casado, M., Koponen, T., Ratnasamy, S., Ghodsi, A., Shenker, S.: Software-defined internet architecture: decoupling architecture from infrastructure. In: Proceedings of the 11th ACM Workshop on Hot Topics in Networks, pp. 43–48 (2012)
21. Ren, W., et al.: Privacy-preserving using homomorphic encryption in mobile IoT systems. Comput. Commun. **165**, 105–111 (2021)
22. Sajjad, M., et al.: Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities. Futur. Gener. Comput. Syst. **108**, 995–1007 (2020)
23. Wang, Y., Liang, X., Hei, X., Ji, W., Zhu, L.: Deep learning data privacy protection based on homomorphic encryption in AIoT. Mob. Inf. Syst. **2021**, 1–11 (2021)
24. Xiaoqiang, Z., Mengmeng, W., Guiliang, Z.: Research on the new development of image encryption algorithms. Comput. Eng. Sci. **34**(5), 1–6 (2012)
25. Xu, J., Zhao, B., Wu, Z.: Research on color image encryption algorithm based on bit-plane and chen chaotic system. Entropy **24**(2), 186 (2022)
26. Zhu, H., Peng, Y., Xu, H., Tong, F., Jiang, X.Q., Mirza, M.M.: Secrecy enhancement for SSK-based communications in wireless sensing systems. IEEE Sens. J. **22**(18), 18192–18201 (2022)

Fig. 8. Data of the encrypted AI model file, garbled and unreadable (partial display)

5 Conclusions

In this paper, a model privacy protection solution is proposed to help protect the privacy of model owners and data owners in complex IoT application environments. The improved RC6-plus, JNI middleware approach can more effectively protect model privacy in the IoT environment. Experimental results demonstrate the reasonableness and effectiveness of the approach. Considering AI model applications are increasingly used in new smart IoT software, hardware application products, and other services, and the arithmetic performance of cloud-based servers far exceeds the hardware configuration on the IoT side, this research can be further extended to model encryption and decryption of cloud-based server applications and data privacy protection during remote data transmission of IoT.

References

1. Ahmed, H.E.D.H., Kalash, H.M., Allah, O.F.: Encryption efficiency analysis and security evaluation of RC6 block cipher for digital images. In: 2007 International Conference on Electrical Engineering, pp. 1–7. IEEE (2007)
2. Alenezi, M.N., Alabdulrazzaq, H., Mohammad, N.Q.: Symmetric encryption algorithms: review and evaluation study. Int. J. Commun. Netw. Inf. Secur. **12**(2), 256–272 (2020)
3. Ashraf, A.M., Elmedany, W.M., Sharif, M.S.: Secure IoT data transmission at physical layer using rc6 encryption technique. In: 2022 9th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 307–315. IEEE (2022)
4. Baños-Gonzalez, V., Afaqui, M.S., Lopez-Aguilera, E., Garcia-Villegas, E.: IEEE 802.11 ah: a technology to face the IoT challenge. Sensors **16**(11), 1960 (2016)
5. Dong, B., Shi, Q., Yang, Y., Wen, F., Zhang, Z., Lee, C.: Technology evolution from self-powered sensors to AIoT enabled smart homes. Nano Energy **79**, 105414 (2021)
6. Fang, K., Yang, B., Zhu, H., Lin, Z., Wang, Z.: Two-way reliable forwarding strategy of RIS symbiotic communications for vehicular named data networks. IEEE Internet Things J. (2022)

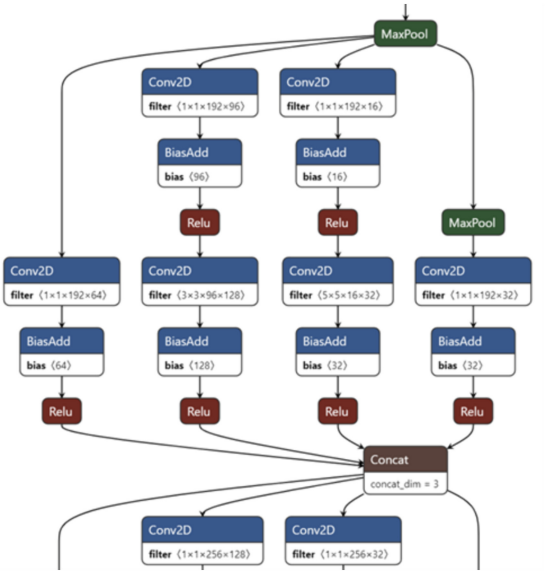


Fig. 7. Structure diagram of AI model before encryption (partial display)

RC6-plus algorithm will use the same algorithmic process but execute the algorithmic steps opposite to the encryption direction. The wheel keys are applied opposite to recover the original message. Therefore, the decryption process of the RC6-plus algorithm takes the same amount of time as the encryption process.

It is important to note that the key length and algorithm parameters used in the RC6-plus encryption and decryption process must be the same to perform the decryption correctly. If a different key length or parameters are used in the decryption phase than in the encryption phase, the decryption process may fail or get an incorrect message.

In addition, this paper also designs the AI-IoT software decoupling architecture so that the API interface layer is decoupled from the JNI layer. The user can only see the outer interface call function but does not know the principle of the internal algorithm [7,9,16], as shown in Fig. 6. The forward inference neural network of the AI model is rewritten using the miniCaffe C++ language, and the source code is obfuscated so that even standard model visualization cracking tools cannot read or write to the model. With the above source code encryption process, the project source code can be compiled to generate .so files, making the system more private, as shown in Fig. 7 and Fig. 8.

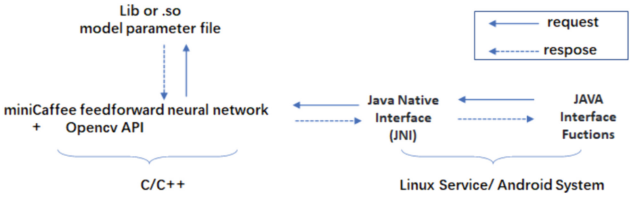


Fig. 6. AI-IOT software decoupling architecture

When the Key length equals 64 bits, the Digital dataset RC6 average encryption time equals 127 ms, while the Digital dataset - RC6-plus average encryption time only takes 89 ms; when the Key length increases to 208 bits, the Digital dataset RC6 average encryption time equals 407 ms, while the Digital dataset - RC6-plus average encryption time only takes 323 ms. When the Key length increases to 208 bits, the Digital dataset RC6 average encryption time equals 407 ms, while the Digital dataset - RC6-plus average encryption time is only 323 ms.

When the key length is 64 bits, the average encryption time of an image data set using the RC6 algorithm is 1896 milliseconds, while the average encryption time using the RC6-plus algorithm is only 1186 milliseconds, and the encryption time of the RC6-plus algorithm is significantly less than that of the RC6 algorithm; when the key length is increased to 208 bits, the advantage of RC6-plus algorithm is more prominent, and its average encryption time is 5137 ms, which takes 299 ms less than the RC6 algorithm. For audio data sets, the average encryption time of the RC6-plus algorithm is only 191 milliseconds when the key length is 64 bits. For digital data sets, the average encryption time of the RC6-plus algorithm is only 464 milliseconds when the key length is 208 bits.

From the results of the analysis of the above experimental data, we draw several conclusions:

1. The encryption time increases as the value of the Key length increases;
2. The encryption time will vary depending on the complexity of the data;
3. The RC6-plus algorithm can be used on text datasets, digital datasets, image datasets, or audio datasets. The encryption performance of the algorithms on the audio data set is significantly better than RC6.
4. These data results show that the RC6-plus algorithm has a shorter encryption time than the RC6 algorithm for different data sets and critical lengths. The advantage is undeniable for longer key lengths.

The experimental data of encryption strength and encryption time of the RC6-plus algorithm do not include the experimental data of decryption of the RC6-plus algorithm. In practical applications, the decryption time and strength are usually related to factors such as the length of the key used for encryption, the data set, and the algorithm version. Usually, the decryption process of RC6-plus is similar to the encryption process, but the order of the keys is reversed. Therefore, the same key and algorithm parameters should be used in the decryption phase as in the encryption phase. In the decryption process, the

Table 2 shows the experimental data of the performance tests on RC6 and RC6-plus encryption algorithms for text and numeric datasets, respectively, and Fig. 4 shows the visual line graph corresponding to Table 2. The data are recorded in the table.

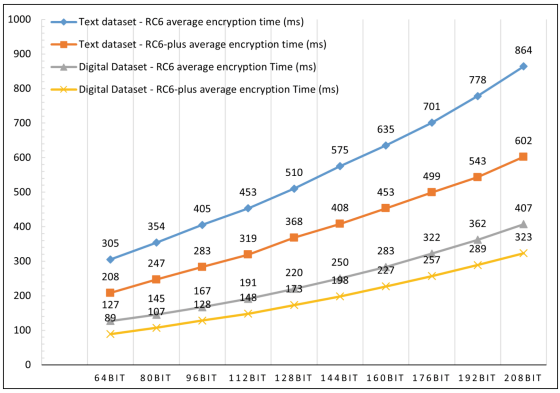


Fig. 4. Visual line chart of performance comparison of encryption algorithms for text and digital datasets

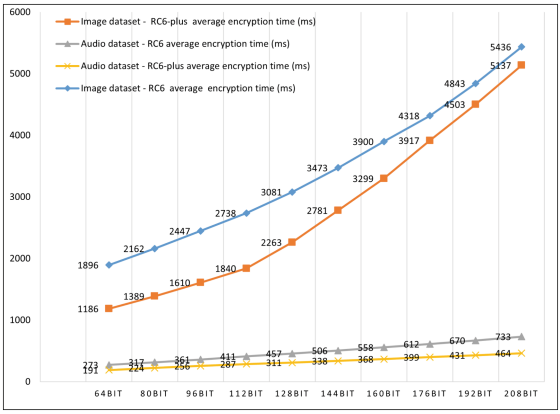


Fig. 5. Performance comparison of image and audio dataset encryption algorithms

When the Key length is equal to 64bit, the Text dataset RC6-plus average encryption time is equal to 305 ms, while the Text dataset RC6-plus average encryption time is only 208 ms, RC6-plus encryption time is less than RC6; When the Key length increases to 208 bits, the advantage of RC6-plus is more apparent, and the average encryption time of RC6-plus is 602 ms, which is 262 ms less than that of RC6.

Table 2. Performance comparison of encryption algorithms for text and digital datasets

Key Length	Average encryption time (ms) of Dataset - Algorithm			
	Text - RC6	Text - RC6-plus	Digital - RC6	Digital - RC6-plus
64bit	305	208	127	89
80bit	354	247	145	107
96bit	405	283	167	128
112bit	453	319	191	148
128bit	510	368	220	173
144bit	575	408	250	198
160bit	635	453	283	227
176bit	701	499	322	257
192bit	778	543	362	289
208bit	864	602	407	323

from web crawlers. In this experiment, 1000 copies of each data type were sampled randomly from these original data sets as the original data for the experiments in this paper. The average encryption time of each data type is taken after the experiment. The test data will be listed in a table, where each row represents a test key length, and each column represents the encryption time and encryption strength of each test item in the dataset.

In this paper, we experimentally compare the encryption time and strength of the traditional RC6 algorithm and RC6-plus algorithm on text, digital, image, and audio datasets, as shown in Table 2, Fig. 4, Table 3, and Fig. 5.

Table 3. Performance comparison of encryption algorithms for image and audio datasets

Key Length	Average encryption time (ms) of Dataset - Algorithm			
	Image - RC6	Image - RC6-plus	Audio - RC6	Audio - RC6-plus
64bit	1896	1186	273	191
80bit	2162	1389	317	224
96bit	2447	1610	361	256
112bit	2738	1840	411	287
128bit	3081	2263	457	311
144bit	3473	2781	506	338
160bit	3900	3299	558	368
176bit	4318	3917	612	399
192bit	4843	4503	670	431
208bit	5436	5137	733	464

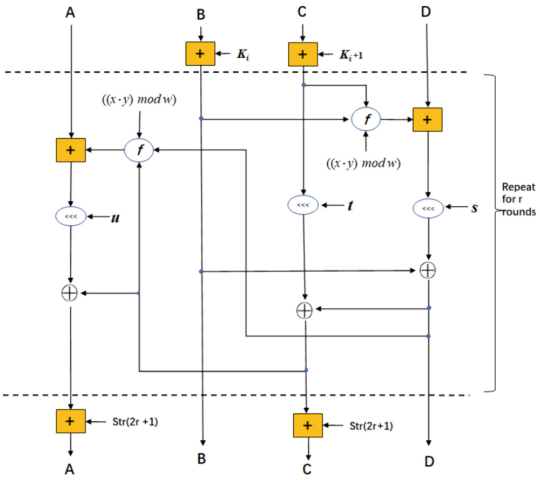


Fig. 3. Principle of wheel function of RC6-plus algorithm

Table 1. Raspberry Pi 4 Model B configuration parameters

Configuration items	Specification
CPU	Broadcom BCM2711, quad-core Cortex-A72(ARM v8), 64-bit SoC at 1.5 GHz
Memory	LPDDR4 SDRAM, 4 GB, MicroSD card slot
Network	Gigabit Ethernet, dual-band 802.11ac wireless card (with optional Bluetooth 5.0)
USB	2 * USB 3.0 and 2 * USB 2.0 ports
Video/Audio Output	2 * micro-HDMI ports supporting up to 4K resolution
Audio	Stereo output, stereo MIC, support Bluetooth audio output
GPIO	40 * GPIO pins
Operating System	Raspberry Pi OS (Debian-based operating system)

4 Experimental Results and Analysis

In this work, we tested the performance of the proposed scheme on a real Raspberry Pi-based IoT device. To evaluate the performance of the proposed scheme, we exclude the time consumed on the communication channel as it heavily depends on the network traffic. The experimental setup focuses only on the performance tests for the time used for encryption, RC6-plus operation, and decryption. The RC6-plus instances are all running on the latest Raspberry Pi (Raspberry Pi 4 Model B) with the system parameters configured, as shown in Table 1.

The following experiments are based on improved RC6 cryptographic algorithms with different bit cell lengths (64bit 208bit), using six datasets provided, each testing ten RC6 algorithms with different key lengths. The datasets are derived from the homebrew program Automatic Random Sequence, COCO dataset, ImageNet, CIFAR, MNIST, PASCAL VOC, SQuAD, Labeled Faces in the Wild, UCI Machine Learning Library, and with the addition of some data

and 4 round constants K_i in each round, and 4 outputs A', B', C', D', then the computation process of the round function can be expressed as

$$B \leftarrow B + K_i \tag{15}$$

Pass the new value B calculated in Eq. 1 into the f function:

$$D \leftarrow D + f(B, C) \tag{16}$$

$$D \leftarrow D <<< s \tag{17}$$

$$D \leftarrow D \oplus B \tag{18}$$

The new value D is calculated and rounded with C:

$$C \leftarrow C + K_i + 1 \tag{19}$$

$$C \leftarrow C <<< t \tag{20}$$

$$C \leftarrow C \oplus D \tag{21}$$

Immediately afterward, the new values C and D are rounded with A:

$$A \leftarrow A + f(C, D) \tag{22}$$

$$A \leftarrow A <<< u \tag{23}$$

$$A \leftarrow A \oplus C \tag{24}$$

After following the above cryptographic round, a new set of output values is obtained:

$$A' \leftarrow A, B' \leftarrow B, C' \leftarrow C, D' \leftarrow D \tag{25}$$

where $i = 1, 2, ..., r, s, t, u$ are the computed parameters, and f(x, y) denotes the improved RC6-plus algorithm. A nonlinear function is introduced in the program for converting inputs to outputs. Specifically, the function $f(x, y)$ can be defined as

$$f(x, y) = (x \oplus y) <<< ((x \cdot y) \mod w) \tag{26}$$

This function is a nonlinear function that converts the inputs B and C into an output word D. The class Similarly, another nonlinear function $f'(x, y)$ can be defined as

$$f'(x, y) = (x \oplus y) <<< (w - ((x \cdot y) \mod w) \tag{27}$$

This function converts inputs C and D into an output word A to further disrupt data flow during encryption. The above is the basic structure schematic of the wheel function in the RC6-plus algorithm, in which more complex nonlinear functions are introduced to enhance the security of the encryption algorithm. In practical applications, the wheel function can be adjusted and optimized according to specific needs to improve the strength and performance of the encryption algorithm. RC6-plus is similar to encryption and decryption, which is not repeated here in this paper, and the wheel function of the RC6-plus algorithm can be viewed as shown in Fig. 3.

The detailed process of RC6-plus algorithm improvement has the following steps:

1. Adjust the number of rounds r and w-bit word length to balance security and algorithm performance
   First, we assume the original RC6 algorithm uses r-rounds and w-bit word lengths. For each word A, B, C, D, we can calculate its output using the following equation:

$$A' = ((A \oplus B) <<< s) \oplus K_0 \tag{9}$$

$$B' = ((B \oplus C) <<< t) \oplus K_1 \tag{10}$$

$$C' = ((C \oplus D) <<< u) \oplus K_2 \tag{11}$$

$$D' = ((D \oplus A) <<< v) \oplus K_3 \tag{12}$$

   where $\oplus$ denotes the XOR operation, $<<<$ denotes a circular shift, K_i is a round constant, and s, t, u, and v are parameters that need to be calculated based on the w-bit word length. The formulae for these parameters are as follows:

$$w = 32 \rightarrow r = 20, s = 7, t = 2, u = 13, v = 8 \tag{13}$$

$$w = 64 \rightarrow r = 20, s = 35, t = 5, u = 31, v = 16 \tag{14}$$

   Suppose we want to improve the algorithm to increase its performance. In that case, we can reduce the number of rounds r or decrease the w-bit word length to reduce the amount of computation for encryption. However, this will also reduce the security of the algorithm.
2. Modify the generation method of the RC6-plus algorithm wheel constant K_i
   Increase the randomness and complexity of wheel constant generation to enhance the strength of the encryption algorithm. The wheel constant Ki of the RC6 algorithm is derived from a specific key. If this key can be guessed or leaked, then the security of the encryption is threatened. Therefore, we need to enhance the randomness and complexity of the wheel constant generation to improve the strength of the algorithm. We use more complex key derivation algorithms or introduce more wheel constants to increase the randomness and complexity of encryption.
3. Optimization of operations in the encryption wheel
   The original RC6 algorithm uses relatively simple arithmetic, so we must introduce more complex nonlinear functions and improve the algorithm using iso-or and circular shifts. Simple attack methods can break this simple arithmetic, so we can optimize the arithmetic in the encryption wheel by introducing more complex nonlinear functions to increase the strength of the algorithm.
   With the above three improvements, we can improve the security and performance of the RC6 algorithm to make it more suitable for practical applications, and the improved RC6 algorithm is noted as RC6-plus. Suppose the RC6-plus algorithm uses r rounds of encryption, with 4 inputs A, B, C, D,

2. In the algorithm of RC6, Str(i) represents the subkey word, while "<<<" and ">>>" represent the controlled left rotation and right rotation, respectively. The symbol controls the amount of rotation, followed by the number of rotations is controlled by the lowest 5 bits of the number following the symbol. In addition, to meet the fixed grouping bit length requirement, RC6 also uses a quadratic function B*(2B+1) to strengthen the diffusion property, which is very different from most other encryption and decryption algorithms. The graphical representation of the wheel function is shown below (in Fig. 2), where $f(x)$ represents the following nonlinear invertible function:

$$f(x) = (x * (2x + 1)) <<< lg(w) \tag{8}$$

3. RC6 is a high-performance, highly flexible group iterative cipher whose compact and transparent architecture makes it widely used in monolithic microcontrollers. In addition, RC6 performs even better in application scenarios such as fingerprint recognition and POS machines. The data-dependent cyclic nature of RC6 can significantly improve encryption efficiency while its memory requirements are relatively low, and the highly integrated internal cache technology can significantly reduce production costs.
Although the RC6 algorithm is designed for simplicity and efficiency, it still has some shortcomings, such as the lack of performance of the nonlinear function f because the bit diffusion of f is unidirectional. The diffusion speed is slow, and the average computation is w/2 = 16 additions due to the use of multiplication in f, so the nonlinear function becomes the bottleneck of the operation speed. In addition, RC6 has significant differences in the encryption and decryption algorithms, which is also a drawback. To address these issues, some improvements were made to RC6 as follows.

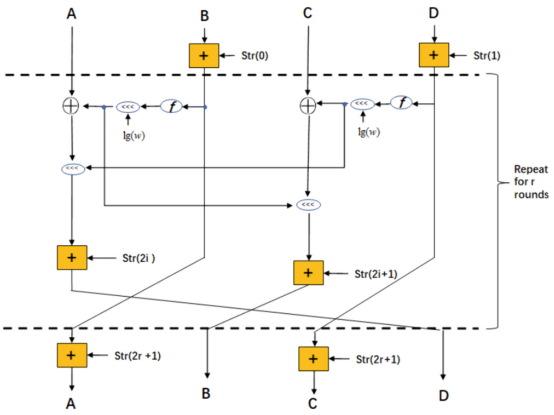


Fig. 2. Principle of RC6 algorithm wheel function

transmitted to RC6-plus for encryption and decryption via JNI. If the secret key is decrypted successfully, the gate control valve (GS, Gate Switch) is opened, allowing the data received by the I/O to reach the feedforward neural network layer through the JNI layer to complete the inference, which is often accompanied by the process of model loading. The final result is again transmitted to the terminal display device through the I/O interface.

3.2 Key Algorithm Study

In this section, we analyze the traditional RC6 algorithm in detail and find that RC6 has certain defects in the application of this project [7,9,22]. Based on this, we propose the RC6-plus encryption algorithm with flexible control of the number of round bits and successfully apply it to our intelligent IoT terminal AI project, achieving good results.

The detailed process of the RC6 algorithm has the following steps:

1. RC6 is one of the AES candidate algorithms. It is an improved version of the RC5 algorithm. (w, r, b) in RC6-$w/r/b$ denotes the operation word length, the number of iteration rounds, and the user master key length, respectively. Usually, we choose the arithmetic word length w = 32 bits (bit). The plaintext packet length is 4 characters (128 bits). RC6 consists of input encryption, r rounds of iteration, and output transformation.
   Input encryption:

$$(A, B, C, D) = (A, B + Str(0), C, D + Str(1)) \tag{1}$$

   Round r iteration:

$$t = [B * (2B + 1)] <<< lg(w); \tag{2}$$

$$u = [D * (2D + 1)] <<< lg(w); \tag{3}$$

$$A = [A \oplus t <<< u] + Str(2i); \tag{4}$$

$$C = [C \oplus u <<< t] + Str(2i + 1); \tag{5}$$

$$(A, B, C, D) = (B, C, D, A); \tag{6}$$

   Output transformation:

$$(A, B, C, D) = (A + Str(2r + 2), B, C + Str(2r + 3), D); \tag{7}$$

2. A new improved RC6 (Rivest cipher 6) algorithm is proposed for enhanced model encryption, denoted as RC6-plus;
3. An improved design scheme between AI-IOT software layers is proposed to further enhance security.

3 Proposed Method

This section focuses on two parts; the first part is an explanation of the model training steps and the architecture diagram of the AI-IOT model inference system; the second part introduces the traditional RC6 while proposing the RC6-plus algorithm and its improvement process; the reader will learn about the method of AI model security protection research for IoT terminals described in this paper.

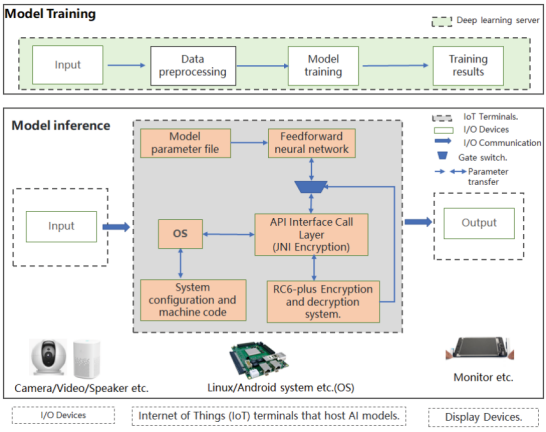


Fig. 1. Server model training and AI-IOT model inference system architecture

3.1 System Architecture Design

As the architecture is shown in Fig. 1, the model training task is done on the deep learning GPU server, the model inference is made on the smart IoT terminal, and the model can run on Linux or Android operating systems. The basic steps of model training (feedback neural network) include data input, data preprocessing, feedback neural network calculation, and parameter storage. Among them, model inference (feedforward neural network) mainly includes input, model parameter loading, and model computation inference. The data on the input side can come from the camera, microphone, or locally stored data, and the output receiver can be the display or stored in the database. This paper focuses on the software system architecture approach to the smart IoT terminal. The operating system (OS, Operate System) extract is coded by the connected terminal machine and

3. Encryption algorithm research
   Encryption algorithms protect secure communication between IoT endpoints and AI models. Researchers have recently researched encryption algorithms to ensure data security during the communication process. For example, researchers have proposed trusted authentication techniques, secure transmission, and data encryption algorithms [2,3,7,15,18,19,24,25].
4. Multi-level security mechanism design
   In order to enhance the security between IoT terminals and AI models, the researchers proposed a multi-level security mechanism design. This approach will strengthen security in several aspects, such as model management, model usage, and model storage, to maximize the security of the IoT terminal system [2,7,9,11,12,14,16–18].

The above are only some of the research contents related to the research of AI model security protection methods for IoT terminals. However, the research in this area is still in its initial stage, and further in-depth research and exploration are needed in many aspects. Therefore, it is of significant theoretical and application value to research the AI model security protection method for IoT terminals [7,11–13,19,24].

Based on this, this research will draw on domestic and international research results and methods to explore a more comprehensive, detailed, and feasible AI model security protection method from various aspects, such as data security, model protection, and encryption algorithms. The research aims to solve the challenges faced by AI model security on IoT endpoints, protect the security and privacy of devices and data, and provide useful academic and practical references for the innovative development of related fields [2,3,7,18,25].

This paper first introduces the basic concepts of IoT and AI models and analyzes the existing AI model security threats and the limitations of existing solutions. Secondly, the paper also proposes a security protection method based on encryption and access control and details the implementation process and related technical details of the method.

Finally, the paper verifies the effectiveness and feasibility of the proposed method through experiments. The experimental results show that the method has high security and scalability and can provide more comprehensive protection for AI models in IoT applications.

Overall, the application of AI-IoT offline identification devices can effectively improve the efficiency of data sampling, reduce the cost of manual work, enhance accuracy, and is suitable for data collection and processing tasks in various complex environments. The development of this technology plays an essential role in developing the modern manufacturing industry, promoting industrial upgrading, and enhancing national strength. The research results of this paper are of great significance to the development and popularization of IoT applications and provide a helpful reference for AI safety.

The main contributions of this work can be summarized as:

1. An efficient system architecture of AI model is proposed for IoT terminals;

Internet of Things (AI-IOT) offline identification device can realize real-time processing and identification of data by deploying the model on the body of the device, and the data processing method that does not depend on the transmission environment reduces the risk of data transmission interruption and solves the problem of its online real-time identification instability, which has high practical value. In archaeology, exploration, and geological examination, AI-IOT can process and comprehensively analyze a large amount of data. Based on the characteristics of offline processing technology, this AI-IOT can also work autonomously through unitized design and intelligent scheduling technologies to further improve work efficiency. Conventional monitoring systems may have signal interruptions and analysis errors in bad weather, such as rain, wind, and lightning. At the same time, AI-IOT can deploy AI models on the equipment to realize real-time monitoring and grasp the comprehensive situation of the machine, environment, and other parameters, and conduct intelligent analysis and processing of this, increasing the application areas of artificial intelligence [8,21]. At the same time, the application of AI-IOT is becoming increasingly widespread and will face many challenges, including physical security, identity identification issues, external attacks, vulnerability issues, data validation, model security, program update mechanism, and communication security. In particular, AI models involve a large amount of data and privacy; once attacked and maliciously changed, it is easy to affect the stability and reliability of the whole system, and people start to pay attention to the impact of read and write operations of smart IoT terminal device data on the privacy and security of AI models. Therefore, how to protect the AI models of IoT terminals has become a critical research direction [1,20,21,23].

This paper is organized as follows: Sect. 2 reviews the current research status of AI-IOT; Sect. 3 reviews an efficient AI model application system architecture for IoT terminals and crucial algorithm research (RC6, RC6-plus); Sect. 4 focuses on the research content experimental results and analysis process. Finally, conclusions are drawn in Sect. 5.

2 Related Work

At present, AI model security protection for IoT terminals in academia and industry is actively carrying out relevant research, mainly around the following aspects:

1. Data security protection
   For the security threats and risks faced by the data security of IoT terminals, researchers have proposed a series of data security protection techniques. For example, encryption technology is used for secure data storage, integrity protection, and access control mechanisms [6,8,10,21,23].
2. AI model protection
   In response to the security risks faced by AI models, researchers have developed a series of protection techniques for AI models. For example, AI models are encrypted, compressed, and cut to increase the security of the models; a reliable AI algorithm framework is used to ensure the reliability of model training and inference [1,13,20,23,26].

Artificial Intelligence Model Based Security Protection Method for IoT Applications

Xiaolong Luo[1], Xiaoli Chen[2(✉)], Jie Wei[1], Liang Zhang[2], Luping Xu[2], and Bijun Zhao[2]

[1] Zhejiang Water Conservancy Information Publicity Center, Hangzhou 310009, China
[2] Zhejiang Ponshine Information Technology Co., Ltd., Hangzhou 311100, China
chenxiaoli@ponshine.com

Abstract. The issue of model privacy security is increasingly affecting the application systems of Artificial Intelligence Internet of Things (AI-IOT) terminals, where it is challenging to protect the privacy of the underlying AI models. In this paper, we propose a security protection RC6-plus algorithm based on cryptography and access control for AI model security in IoT applications. Specifically, the proposed method effectively protects the privacy of crucial algorithms in the program by encrypted storing the model parameters, as well as storing and code obfuscating the neural network structure and parameters of the AI model independently while adding the isolation treatment of the JNI communication layer. The results of the experiments verify the effectiveness of the proposed method.

Keywords: Model privacy security · internet of things · artificial intelligence model · encryption technology · access control

1 Introduction

With the continuous development of society and technology, Internet of Things (IoT) technology has been widely used in various fields. The IoT is gradually becoming an indispensable part of people's life and work because of its intelligence and connectivity. In recent years, IoT and Artificial Intelligence (AI) technologies are becoming more and more closely integrated, and more and more AI technologies are being applied to IOT end devices. Such devices also have the ability of offline recognition, which can realize various applications in highland, deep sea, remote areas, archaeology, exploration, and geological examination [4,5,10].

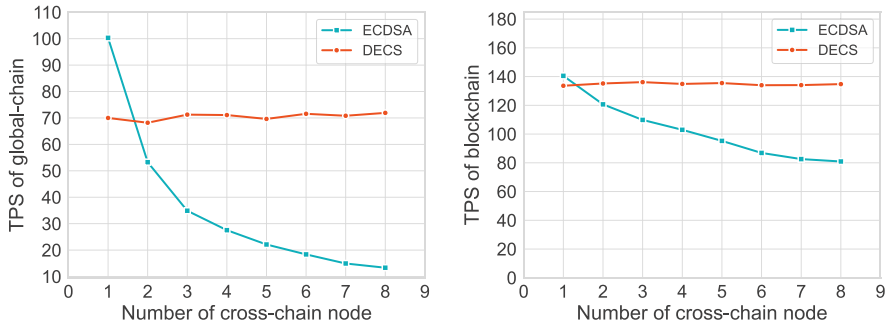
In the plateau, deep sea, and other particular environments, conventional networking equipment may have the problem of unstable network transmission, leading to interruption in the data transmission process. The Artificial Intelligence

X. Chen—This work was funded by the Zhejiang Provincial Department of Water Resources Science and Technology Plan Project, Zhejiang, China (Project no. RC2238).

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 143–157, 2024.
https://doi.org/10.1007/978-3-031-65126-7_15

16. Zhang, H., Lao, L., Shu, C., Xiao, B.: Analysis of the communication traffic model for permissioned blockchain based on proof-of-work. In: IEEE International Conference on Communications (ICC 2021). IEEE International Conference on Communications, IEEE (2021). https://doi.org/10.1109/ICC42927.2021.9500333, Telus; Huawei; Ciena; Nokia; Samsung; Qualcomm; Cisco; Google Cloud, iEEE International Conference on Communications (ICC), ELECTR NETWORK, JUN 14-23, 2021
17. Zhou, J., Feng, G., Wang, Y.: Optimal deployment mechanism of blockchain in resource-constrained IoT systems. IEEE Internet Things J. **9**(11), 8168–8177 (2022). https://doi.org/10.1109/JIOT.2021.3106355

5. Dai, H.N., Zheng, Z., Zhang, Y.: Blockchain for internet of things: a survey. IEEE Internet Things J. **6**(5, SI), 8076–8094 (2019). https://doi.org/10.1109/JIOT.2019.2920987
6. Deng, L., Chen, H., Zeng, J., Zhang, L.-J.: Research on cross-chain technology based on sidechain and hash-locking. In: Liu, S., Tekinerdogan, B., Aoyama, M., Zhang, L.-J. (eds.) EDGE 2018. LNCS, vol. 10973, pp. 144–151. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94340-4_12
7. Genc, Y., Afacan, E.: Design and implementation of an efficient elliptic curve digital signature algorithm (ecdsa). In: Chakrabarti, S., Paul, R., Gill, B., Gangopadhyay, M., Poddar, S. (eds.) 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1026–1031 (2021). https://doi.org/10.1109/IEMTRONICS52119.2021.9422589. IEEE; Inst Engn & Management; IEEE Vancouver Sect; IEEE Toronto Sect; SMART; Univ Engn & Management , iEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), ELECTR NETWORK, APR 21-24, 2021
8. Ghosh, B.C., Bhartia, T., Addya, S.K., Chakraborty, S.: Leveraging public-private blockchain interoperability for closed consortium interfacing. In: IEEE INFOCOM 2021 - IEEE Conference on Computer Communications, pp. 1–10 (2021). https://doi.org/10.1109/INFOCOM42981.2021.9488683
9. Hope-Bailie, A., Thomas, S.: Interledger: Creating a standard for payments. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 281–282. WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2016). https://doi.org/10.1145/2872518.2889307
10. Kang, E.S., Pee, S.J., Song, J.G., Jang, J.W.: A blockchain-based energy trading platform for smart homes in a microgrid. In: Proceedings of 2018 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 472–476 (2018). IEEE; Nagoya Inst Technol, 3rd International Conference on Computer and Communication Systems (ICCCS), Nagoya, JAPAN, APR 27-30, 2018
11. Lu, Q., Xu, X.: Adaptable blockchain-based systems a case study for product traceability. IEEE Softw. **34**(6), 21–27 (2017). https://doi.org/10.1109/MS.2017.4121227
12. Makhdoom, I., Abolhasan, M., Abbas, H., Ni, W.: Blockchain's adoption in IoT: the challenges, and a way forward. J. Netw. Comput. Appl. **125**, 251–279 (2019). https://doi.org/10.1016/j.jnca.2018.10.019
13. Sun, Y., Yi, L., Duan, L., Wang, W.: A decentralized cross-chain service protocol based on notary schemes and hash-locking. In: 2022 IEEE International Conference on Services Computing (SCC), pp. 152–157 (2022). https://doi.org/10.1109/SCC55611.2022.00033
14. Syta, E., et al.: Keeping authorities honest or bust with decentralized witness cosigning. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 526–545 (2016). https://doi.org/10.1109/SP.2016.38
15. Tschorsch, F., Scheuermann, B.: Bitcoin and beyond: a technical survey on decentralized digital currencies. IEEE Commun. Surv. Tutor. **18**(3), 2084–2123 (2016). https://doi.org/10.1109/COMST.2016.2535718



(a) TPS of global-chain vs. Number of cross-chain nodes.

(b) TPS of entire blockchain vs. Number of cross-chain nodes.

Fig. 5. Blockchain performance comparison.

6 Conclusion

In this paper, we present a novel decentralized and efficient cross-chain scheme tailored for IoT systems. Our constructed system demonstrates excellent adaptability to the diverse and intricate nature of IoT environments. Furthermore, our proposed cross-chain solution effectively addresses the pervasive issue of excessive centralization within current cross-chain technologies. The incorporation of BLS signature-based data verification alleviates the burden of high storage requirements associated with the ECDSA scheme. Experimental results show that our scheme has better decentralization and efficiency, and the blockchain throughput has a good performance. We hope our scheme can be used as a high-performance tool for IoT systems to provide efficient, secure protection for IoT data.

References

1. Ahluwalia, S., Mahto, V, R., Guerrero, M.: Blockchain technology and startup financing: a transaction cost economics perspective. Technol. Forecast. Soc. Change **151** (2020). https://doi.org/10.1016/j.techfore.2019.119854
2. Back, A., et al.: Enabling blockchain innovations with pegged sidechains. **72**, 201–224 (2014)
3. Boudguiga, A., et al.: Towards better availability and accountability for IoT updates by means of a blockchain. In: 2017 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 50–58 (2017). https://doi.org/10.1109/EuroSPW.2017.50
4. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. IEEE Access **4**, 2292–2303 (2016). https://doi.org/10.1109/ACCESS.2016.2566339

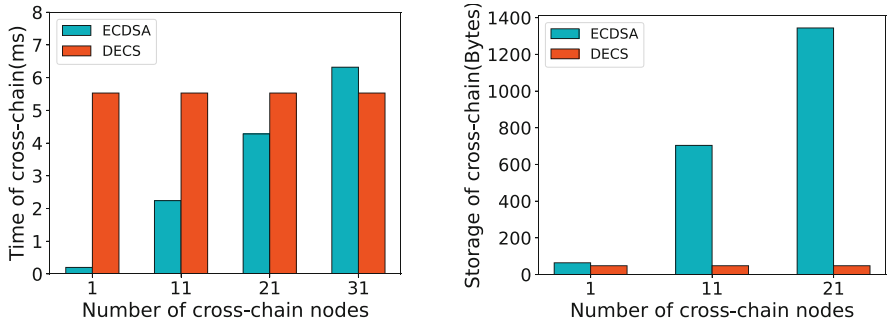
Table 1. Time cost for each step of the BLS signature.

Notation	Description	execution time
T_z	Take a random number in T_z	0.0183 ms
T_{G_1}	Point multiplication in G_1	0.5482 ms
T_{G_2}	Point multiplication in G_2	0.6671 ms
T_a	Point addition in G_1	0.3030 ms
T_p	Polynomial calculation time	0.0053 ms
T_h	Hashing with SHA256	0.0027 ms
T_e	Verification of Bilinear Functions in T_z	3.9822 ms

Figure 4a shows the time cost required by the two schemes. The experiment starts with the most basic notary scheme, which means that only 1 node is involved in the cross-chain. We can see that at this point the ECDSA scheme significantly outperforms the DECS scheme. However, as the number of cross-chain nodes increases, the time spent by the ECDSA scheme keeps increasing. This is because the ECDSA scheme is unable to aggregate signatures. The verifier needs to verify all the signatures one after another. On the other hand, the DECS scheme can keep the verification time at 5.53 milliseconds due to the aggregation of signatures.

Figure 4b illustrates the storage cost required by both schemes. Similar to the time cost. The ECDSA scheme has increasing storage space as the number of cross-chain nodes. In contrast, the DECS scheme only has one MPK for each cross-chain process, regardless of the number of cross-chain nodes. So the storage cost remains constant. Since this scheme uploads the public key into the global-chain, the storage size will significantly affect the overall performance of the blockchain. This gives the DECS scheme a more obvious advantage.

Blockchain Performance: Figure 5a and Fig. 5b illustrates the TPS of the blockchain under both scheme. Where Fig. 5a shows the TPS of the global-chain. As the number of cross-chain nodes increases. The data size that needs to be stored on the global-chain increases for the ECDSA scheme. The throughput of the blockchain also keeps decreasing. The same effect is seen in Fig. 5b. Figure 5b shows the test performed for the entire blockchain network, which includes two local-chains and one global-chain. As we can be seen from the figure, the performance of the ECDSA scheme keeps decreasing. While the performance of the DECS scheme remains stable in both experiments. When the number of cross-chain nodes reaches 8, the DECS scheme significantly outperforms the ECDSA scheme.



(a) Time cost of cross-chain vs. number of closs-chain nodes

(b) Storage cost of cross-chain vs. number of closs-chain nodes

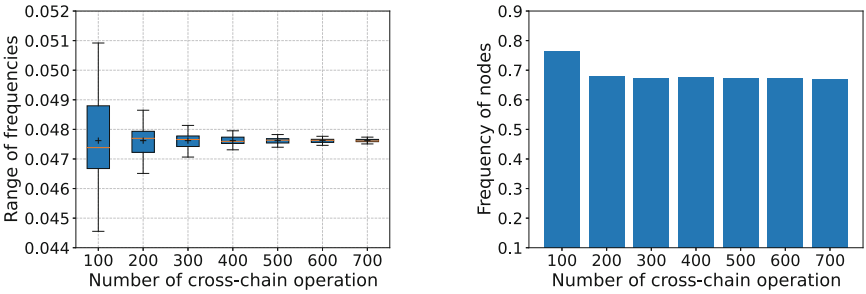
Fig. 4. Time and Storage cost of cross-chain.

the local-chains has 9 nodes and the other has 12 nodes. The two local-chains are independent of each other. The blockchain network adopts the Raft consensus, with a maximum blocking time of 7 s, a maximum number of 100 transactions, and a maximum block size of 100M. At the same time, the experiment used the mainstream testing tool Caliper of the Hyperledger Fabric to test blockchain performance.

Frequency of Nodes Participating in Cross-chain: The first experiment targets the Reputation-Based Node Selection Mechanism to test whether the selection of cross-chain nodes is decentralized. We initiate a cross-chain request to a local-chain of 12 nodes. And set the number of cross-chain nodes to 8, and the reputation parameter $\alpha = 0.5$, to test the frequency of each node participating in the cross-chain.

The experimental results are shown in Figs. 3a and Figs. 3b. Figures 3a is a box plot of the frequency of all nodes participating in cross-chaining. It can be seen that there is still an obvious gap in the frequency of the nodes when completing 100 cross-chaining. And as the number of cross-chaining increases. The gap in frequency gradually decreases. Each node can be selected evenly. Figures 3b count the 8 nodes with the lowest network latency. And calculate the sum of their frequencies. It can be seen that at 100 experiments, the frequency is the highest at 0.76. With the increase in the number of experiments, the value gradually decreases and reaches a stable value of 0.67. This number is the ratio of the number of nodes to the number of all nodes. It shows that our scheme is able to decentralize the selection of nodes participating in cross-chain.

Time and Storage Cost of Cross-chain: We compare the DECS proposed in this paper with the ECDSA-based notary scheme [9]. Compare the time and storage cost used to validate cross-chain data for both. First, We performed a step-by-step test of BLS signatures. The experimental results are shown in the Table 1. Each step was experimented with 100 times and averaged.



(a) Range of all node frequencies vs. Number of cross-chain

(b) Frequency of the 8 nodes with the lowest latency vs. Number of cross-chain

Fig. 3. Frequency of nodes participating in cross-chain.

Reputation Update Process: We use smart contracts to automatically update the reputation scores of nodes. The core idea of this contract is to consider the frequency of participation in cross-chaining and network latency of each node, achieving a balance between frequency and network latency. After the node participant in the cross-chain, its called coefficient c_i is updated. The smaller the c_i is, the more likely it is to be selected as a participant for cross-chain. The specific description is as follows:

1. After the blockchain network is initialized, the smart contract awards cross-chain participation status to the nodes with high trustworthiness. And initialize their called coefficients $c_i = c_(init)$.
2. Cross chain initiator s_r is for each strong node v_i Calculate reputation p_i. And sort the obtained values in ascending order, determine the priority of the signature, select nodes in order to form a set N, send (t, n) threshold signature requests, and send unselected messages to nodes outside the set N.
3. After collecting c_i, t_i from these nodes, the smart contract computes the reputation p_i for each node v_i. And it sorts the obtained values in ascending order. And select the nodes in order to combine them into a set N. A (t, n) threshold signature request is sent to the nodes in the set, and an unselected message is sent to nodes outside the set N.
4. Node $v_k(k \in N)$ that receives threshold signature request performs the reputation update algorithm of the Eq. (11). And the unselected nodes $v_l(l \in Vandl \notin N)$ perform the reputation update algorithm of the Eq. (12).

5 Performance Analysis

We tested our scheme on the HyperLedger Fabric, using Intel (R) Xeon (R) Silver 4214 CPU with 256 GB RAM with 16TB hard drive. We built a blockchain network with 21 nodes. It includes two local-chains and one global-chain. One of

If the final calibration determines that $e(\sigma, G) = e(h(M), MPK)$ is equal, then the returned output is true, and the opposite is false.

The above six steps describe the cross-chain process of the BLS-based threshold signature scheme. The scheme utilizes a polynomial to hide the master private key in the BLS signature method and generates the private keys of the participants from it. Then a specific number of individual participant signatures on the data are integrated using an elliptic curve bilinear mapping function. The master signature is recovered using Lagrange interpolation. Validation is then performed to determine the validity of the transaction.

4.2 Reputation-Based Node Selection Mechanism in Cross-Chain

Parameter Settings: Before describing this mechanism, We first define the following.

Definition 1: Assuming that there are m nodes within a localized chain. One of the nodes is denoted as v_i. where $1 \leq i \leq m$. Each node maintains a called coefficient c_i. And the initial value is c_Init, which satisfies the following equation:

$$c_{init} = \frac{1}{m} \tag{8}$$

Definition 2: A node that is not part of the current local-chain is selected as a cross-chain requester. It records the network latency t for each node in $V = \{v_1, v_2, ..., v_m\}$. The time factor r_i is then computed for each v_i. r_i satisfies the following equation:

$$r_i = \frac{t_i}{\sum_{j \in V} t_j} \tag{9}$$

Definition 3: Knowing the called coefficient c_i and the time coefficient r_i of a node v_i, the cross-chain initiator determines the weight parameter $\alpha(\alpha \in [0, 1])$, for which it can compute the prestige value p_i, which satisfies the following equation:

$$p_i = \alpha c_i + (1 - \alpha) r_i \tag{10}$$

Definition 4: After each cross-chain completion, each selected node v_i needs to be updated with the following equation:

$$c_i = c_i + \frac{1}{mn} (i \in N) \tag{11}$$

And the unselected node v_i also needs to update with the following equation:

$$c_i = c_i - \frac{1}{m(m-n)} (i \in V and i \notin N) \tag{12}$$

After the computation is completed, KGC sends the threshold encryption private key to the selected nodes through a secure channel.

Signature: $Sign(SK, m, t, h) \rightarrow \sigma_i$

Each participant i receives the private key and signs the data $m = \{0,1\}^*$. First, participant i hashes the data and maps the resulting hash digest to G_1: $h(M) \in G_1$, and $M = m||d$, with d equal to 0, 1, 2, ... This is because if the value obtained by hashing m is mapped directly onto the curve, there is a 50% probability that it will not map to a particular point. Therefore, the value of d is increased until the point is successfully mapped. The signature σ_i of participant i will be as shown in Eq. (5:)

$$\sigma_i \leftarrow x_i \times h(M) \in G_1 \tag{5}$$

Aggregation: $Aggregate(\sigma) \rightarrow \sigma$

The node A_j with the highest reputation score is responsible for aggregating signatures. It collects signatures from n participants. When it receives a signature group σ generated from the set Q combined by t participants. and the individual signatures within the signature group are verified. A_j can obtain the signature of the master private key MSK on the data based on Lagrange interpolation.

$$\begin{aligned}
\sigma &= \sum_{i \in Q} \sigma_i \prod_{j \in Q, j \neq i} \frac{j}{j-i} \\
&= \sum_{i \in Q} (h(M)) \times x_i \prod_{j \in Q, j \neq i} \frac{j}{j-i} \\
&= h(M) \times \sum_{i \in Q} x_i \prod_{j \in Q, j \neq i} \frac{j}{j-i} \\
&= h(M) \times x
\end{aligned} \tag{6}$$

Response: After signature aggregation, A_j packages the cross-chain data and MSK in Cross-chain.response. Then A_j sends the response data to B_i via A_{Cross}.

Verification: $Verify(MPK, \sigma, h(M), G, e) \rightarrow true|false$

After the B_i obtains the complete signature obtained by the endorsement initiator, the signature can be verified based on the properties of the bilinear function. The specific principle of verification is shown in the following equation:

$$\begin{aligned}
e(\sigma, G) &= e(x \times h(M), G) \\
&= e(h(M), x \times G) \\
&= e(h(M), MPK)
\end{aligned} \tag{7}$$

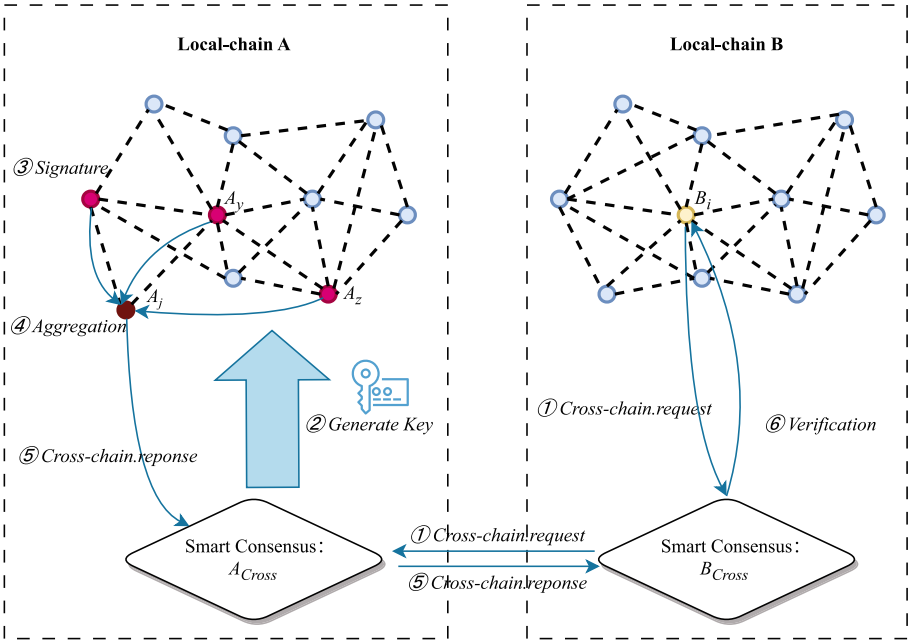


Fig. 2. Cross-chain process of DECS.

SK, and the threshold public key set PK. After the generation is complete, the generation center discloses the master public key MPK and the user group's public key PK. The key generation center randomly selects a random integer as shown in the Eq. (2)

$$x \leftarrow Z_p^* \tag{2}$$

At the same time, KGC set it as the master private key $MSK = x$, and obtains the master public key according to Eq. (3), where G is a randomly selected point from the elliptic curve:

$$v \leftarrow x \times G \in G_2 \tag{3}$$

After this, v will be set as the master public key $MPK = v$. Subsequently, the key generation center randomly selects $t-1$ elements $a_1, a_2, a_3, \ldots, a_{t-1} (a_i \in Z_p^*)$, and set $a_0 = x$, thus constructing a polynomial of order $t-1$:

$$f(x) = \sum_{i=0}^{t-1} a_i x^i \tag{4}$$

Then, KGC calculates x for each participant i $x_i = f(i)$ and set the corresponding threshold private key $sk_i = x_i$. Calculate v_i also according to Eq. (3). And set $pk_i = v_i$. It also calculates for n participants to obtain $SK = x_1, x_2, x_3, \ldots, x_n$ and $PK = \{v_1, v_2, v_3, \ldots, v_n\}$.

4 Cross-Chain Scheme

In this section, we introduce the cross-chain scheme proposed in this paper. It includes a cross-chain interaction process based on BLS signatures and a Reputation-Based Node Selection mechanism.

4.1 Cross-Chain Process Based on BLS Signature

The verification process based on the BLS signature consists of BLS signature algorithm and Shamir based secret sharing algorithm. BLS signature is used to avoid excessive storage cost of the final combined uplink signature, while Shamir secret sharing algorithm is used to achieve a certain number or more node endorsements for specific cross-chain transactions on the basis of BLS signature, thereby ensuring security.

As shown in Fig. 2, our cross-chain process consists of 6 stages: Request, Key Generation, Signature, Aggregation, Response, and Verification. Next, we will describe the 6 stages in detail.

Request: The request stages consists of 2 steps: main steps. First, node B_i of B initiates a request $Cross-chain.request$. Smart contract B_cross receives and parses the request. It will obtain the address of the other party of the cross-chain. After that, it forwards the request to A.

Key Generation: This stage also consists of 2 steps: Setup and Generation.

step1: $Setup(\lambda) \rightarrow \{G_1, G_2, G_T, e, g_1, g_2, p, h\}$

In the setup step, the smart contract B_{Cross} first parses the request. Determine the number of nodes participating in the cross-chain: n. And select the n nodes with the highest reputation. This includes A_j, A_x, A_y, and A_z where A_j is the node with the highest score. After that, B_Cross sends the parameters to the key generation center(KGC).

Then KGC obtains the relevant security parameters p of the algorithm and the elliptic curve bilinear mapping functions $e : G_1 \times G_2 \rightarrow G_T$. Multiplicative Cyclic group G_1, G_2 and its generator g_1, g_2, hash function h for mapping points onto elliptic curves. Specifically, Algorithm Setup(λ) inputs security parameters λ, and outputs as Two multiplicative Cyclic groups of a prime p: G_1, G_2 and a Bilinear map $e : G_1 \times G_2 \rightarrow G_T$. And g_1, g_2 as the generator of G_1, G_2. At the same time, a hash function is used in the scheme to map the hash digest of the signature data to the elliptic curve. These parameters are publicly available in the network:

$$h : \{0,1\}^* \rightarrow G_1 \tag{1}$$

step2: $Generate(G_2, p, t, n) \rightarrow \{MSK, MPK, SK, PK\}$

After initialization of the relevant parameters, KGC generates the master private key MSK, the master public key MPK, the threshold private key set

such as identity information of personal computers, user access information in gateways, etc. Due to the complexity and large number of IoT devices and the heterogeneous nature of IoT data. Adding all IoT devices to the same blockchain will be a very difficult behavior. In this regard, we divide the devices according to their types and organize the IoT devices of the same type to join their respective local-chains.

2. **Local-chain:** The local-chain connects all IoT devices with similar features, such as personal computers and mobile phones that belong to the same smart device. The whole system will have multiple local-chains. It mainly accomplishes the information storage function of the system and is responsible for the data recording, identity granting and world state updating tasks within the local-chains. Taking the first local-chain in Fig. 1 as an example, it is responsible for storing user information, personal wallet, and other data uploaded by cell phones and computers onto the blockchain.
   Moreover, it also records the cross-chain requests and reputation scores of the devices, and these results will be recorded on the local-chain in a summarized form. IoT devices can selectively add one or more local-chains according to their geographic locations and device characteristics.
3. **Global-chain:** There is one and only one global-chain in the system. All blockchain nodes are to be added to the global-chain. The global-chain has two main functions. One is to store the local-chain data abstracts and provide validation function for the local-chain data. When a blockchain node submits data to the local-chain, it can choose to upload the abstracts to the global-chain. Unlike the local-chain which stores a large amount of data, the global-chain only needs to access a small amount of summary information. Second, it provides information records during cross-chain interaction. During cross-chain interaction, some parameters and key information of data exchange will be submitted to the global-chain to ensure security.
4. **Cross-chain module:** The cross-chain module is an important component in linking all local-chains. In this module, we propose a decentralized cross-chain verification scheme. The scheme consists of two parts. The first one is a reputation-based node selection mechanism. We calculate the reputation value of a node based on its network latency and the number of times it has been invoked. Multiple nodes with a good reputation are selected to participate in cross-chain verification. This solves the centralization problem under the notary schemes. It can effectively avoid a single point of failure and improve cross-chain security.
   The second is the data verification technology based on BLS signature. In the cross-chain process, the selected nodes will utilize the BLS signature to sign and verify the data. After that, the node with the highest reputation utilizes BLS to aggregate the signatures of each node to jointly complete the verification of the data. The specific cross-chain process will be introduced in the next section.

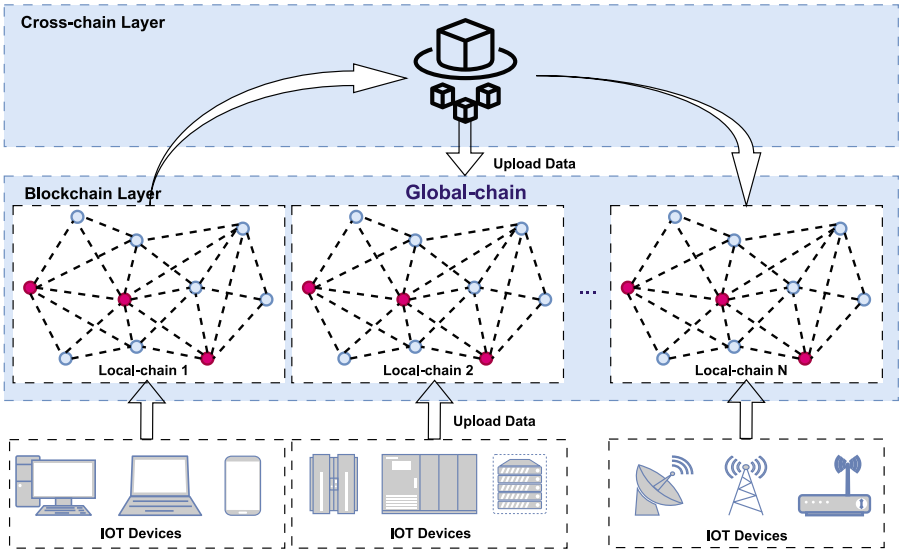


Fig. 1. System model.

which causes it to be more centralized. Once the node itself carries malicious intent or suffers an attack. It will not be able to guarantee the authenticity and trustworthiness of the cross-chain data. However, the cross-chain purpose of this scheme lies in asset transfer, without realizing a universal cross-chain approach.

Sun et al. proposed a decentralized cross-chain scheme [13], which combines a notary mechanism and hash-locking. By setting up multiple notaries and establishing an election mechanism, the degree of centralization of the cross-chain process is reduced. However, the cross-chain purpose of this scheme lies in asset transfer, without realizing a universal cross-chain approach. Ghosh et al. [8] proposed a decentralized gateway architecture that connects private blockchains to end users. The gateway employs a collective signature technique [14] to verify the data. However, this method allows only one-way communication and does not verify the identity of the requester.

3 System Architecture

In this section, we introduce the architecture and components of the system. As shown in Fig. 1, we build a multi-chain network structure in HyperLedger Fabric: including multiple local-chains, a global-chain, and a module that provides cross-chain functionality. IoT devices join the local-chains according to their different features and participate in the whole blockchain system. Next, we will introduce each component of the system in detail.

1. **IoT device:** IoT devices include all networked devices that have a need to participate in the blockchain. They upload important data to the blockchain,

the process of using BLS signatures to verify data in Sect. 4. The experiment results are stated in Sect. 5. Finally, Sect. 6 concludes this article.

2 Related Work

In this section, we introduce the research on blockchain in IoT and cross-chain technology.

2.1 Blockchain in IoT

Existing work applies blockchain to IoT. Blockchain can enhance the security of IoT systems. With encryption and digital signatures by cryptographic keys [7], IoT data can be protected through blockchain. And the smart contract carried by the blockchain can automatically update the firmware of IoT devices and close the vulnerabilities that are susceptible to attacks [4]. Moreover, IoT data stored on the blockchain data can be identified and verified anywhere and anytime. For example, the work of Lu et al. [11] develops a blockchain-based product traceability system that provides traceability to suppliers and retailers. In this way, the quality and originality of products can be checked and verified. Boudguiga et al. [3] proposed a decentralized mechanism to push updates to IoT devices using blockchain. The blockchain is used to record transactions for software updates pushed to the device to prevent malware updates on the device. In this case, there is no need for a trusted agent to deliver the updates as the updates propagated to the device through the blockchain have guaranteed integrity.

The diversity and heterogeneity of IoT devices pose a huge challenge for blockchain [12]. Optimization of blockchain networks is one way to address performance. In 2021, Zhou et al. [17]. proposed an optimization mechanism in resource-constrained IoT systems. They improve the performance of the system by dynamically adjusting optimal block assignments. Zhang et al. [16]. analyze the IoT traffic by establishing a blockchain network that matches the scale of IoT. And they implement a lightweight Bitcoin-like blockchain based on PoW to solve the problem of high traffic load and network congestion.

2.2 Cross-Chain Technology

Cross-chain technology was first applied to the exchange of assets. It is mainly used for the conversion of Bitcoin and Ethereum. In 2014, adam et al. propose sidechain [2], which is a blockchain system independent of bitcoin. Sidechain can access the Bitcoin network and interact with the Bitcoin ledger to enable asset transfers. As a separate blockchain, the technical solutions and consensus mechanisms adopted by sidechain are not restricted by the main chain. The notary schemes [9] is currently the most widely used cross-chain scheme. It set a trusted node in the blockchain system, which is responsible for completing cross-chain operations. However, the notary scheme uses a fixed node for cross-chaining,

during the transaction process [15]. So it leads to a significant gap in blockchain performance compared to centralized systems.

Cross-chain technologies are seen as an effective way to address blockchain scalability and performance. It can join different devices to separate blockchain networks and organize multiple blockchains through cross-chain technologies. The performance of blockchains can be effectively improved. Currently, the mainstream cross-chain technologies include sidechains/relays [2], notary schemes [9], and hash-locking [6]. However, all these technologies have shortcomings. Among them, hash-locking and sidechains/relays are mainly used for asset transfer rather than information interaction. This poses the problem that they focus more on security when crossing chains. And reduce the performance requirements. The notary schemes, on the other hand, suffer from the problem of centralization. In general, the notary mechanism is only responsible by fixed nodes in the cross-chain. The trustworthiness of the cross-chain data is completely guaranteed by the node's own credit. If the node carries malicious intent or it receives an attack, the security of the entire cross-chain process cannot be guaranteed.

In this paper, We propose an decentralized and efficient cross-chain scheme (DECS). First, We construct a multi-chain architecture consisting of multiple local-chains and a global-chain. And we add IoT devices to the local-chain network. Mutually independent local-chains can perform block transactions in parallel. It improves the blockchain and performance. Meanwhile, we propose a scheme in which multiple nodes jointly participate in cross-chain data verification. We design a calculation method for node reputation. It is calculated based on the network latency and invocation frequency of the nodes so that the selection of each node is as average as possible. This reduces the degree of centralization in cross-chain and effectively addresses the shortcomings of the notary schemes. During the cross-chain process, multiple nodes with the highest reputation are selected. They are responsible for verifying the data by using BLS signatures. Our major contributions in this article are summarized as the following aspects:

- We propose an Decentralized and Efficient Cross-chain Scheme in IoT systems. We adopt a multi-chain architecture to adapt the diversity of types of IoT devices and enhance the scalability of the blockchain.
- We design a cross-chain scheme based on BLS signatures and implement a smart contract that automatically updates node reputation. The scheme reduces the degree of centralization of the cross-chain process while guaranteeing the accuracy of cross-chain data.
- We implement and evaluate our scheme on the HyperLedger Fabric, and the results shows that our scheme can effectively improve system performance, and the selection of nodes is uniform and fair in the cross-chain process.

The rest of this paper is structured as follows. In Sect. 2, we introduce the related work of blockchain. In Sect. 3, we described the system architecture and cross-chain interaction mechanisms. And we introduced how to select cross-chain nodes based on reputation. Furthermore, we provide a detailed introduction to

DECS: A Decentralized and Efficient Cross-Chain Scheme in IoT System

Ying Gao(✉), Peihao Zhang, Qiaofeng Pan, and Xianfeng Qiu

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, People's Republic of China
gaoying@scut.edu.cn

Abstract. The rapid development of blockchain technology has demonstrated great potential to revolutionize various application domains, especially in the context of the Internet of Things (IoT). However, the sheer number and diversity of IoT devices pose significant challenges to the scalability and security of blockchain systems. To address these issues, cross-chain technology has emerged as a promising solution. Nevertheless, existing cross-chain solutions suffer from high centralization and inefficiency. Our approach involves constructing a multi-chain network capable of connecting different types of IoT devices, along with designing a collaborative cross-chain mechanism engaging multiple participating nodes. This collective participation diminishes the centralization inherent in the cross-chain process. Specifically, we propose a data verification method based on BLS signature. It aggregates cross-chain data signatures across multiple chains, which leads to a reduced storage burden on the blockchain. Furthermore, we introduce a reputation update algorithm that leverages network latency and cross-chain operation metrics to automatically update node reputation scores via smart contracts. Experimental results demonstrate that our solution achieves better decentralization and efficiency.

Keywords: blockchain · IoT · cross-chain · BLS signature

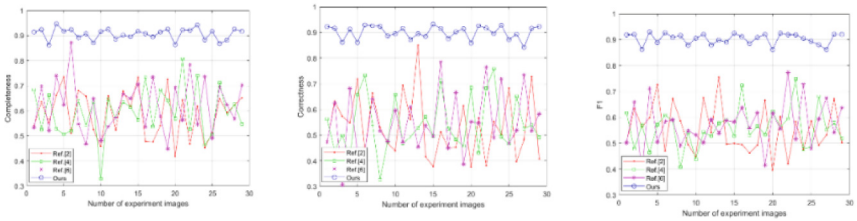
1 Introduction

Blockchain is gradually extending from small-scale applications to multiple fields, showing a bright development prospect. It has been applied to finance [1], food traceability [1], smart home [10], IoT, and other industries. In particular, IoT has a very good application prospect with blockchain due to its own decentralized features [5]. However, another feature of IoT is the large number and diversity of devices. With a large number of devices connected to the blockchain system, higher requirements are placed on the performance of the blockchain system. Unlike centralized systems, blockchain requires all nodes to reach a consensus

This work is supported by the Guangzhou Science and Technology Program key projects (202103010005).

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 128–142, 2024.
https://doi.org/10.1007/978-3-031-65126-7_13

4. Sebasco, N.P., Sevil, H.E.: Graph-based image segmentation for road extraction from post-disaster aerial footage. Drones **6**(11), 315 (2022). https://doi.org/10.3390/drones6110315
5. Vigneshwaran, S.A., Panneer, S.: Situational Analysis of Road Traffic Accidents-Acase of Madurai District rural areas (2020)
6. Zhang, X., Ma, W., Li, C., et al.: Fully convolutional network-based ensemble method for road extraction from aerial images. IEEE Geosci. Remote Sens. Lett. **PP**(99), 1–5 (2019). https://doi.org/10.1109/LGRS.2019.2953523
7. Cheng, G., Wang, Y., Xu, S., et al.: Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. IEEE Trans. Geosci. Remote Sens. **55**(6), 3322–3337 (2017). https://doi.org/10.1109/TGRS.2017.2669341
8. Alamri, A.M.: RoadVec-Net: a new approach for simultaneous road network segmentation and vectorization from aerial and google earth imagery in a complex urban set-up. GISci. Remote Sens. (2021). https://doi.org/10.1080/15481603.2021.1972713
9. Soni, P.K., Rajpal, N., Mehta, R.: Road network extraction using multi-layered filtering and tensor voting from aerial images. Egypt. J. Remote Sens. Space Sci. **24**(2), 211–219 (2021). https://doi.org/10.1016/j.ejrs.2021.01.004
10. Nguyen, T.L., Han, D.: Detection of road surface changes from multi-temporal unmanned aerial vehicle images using a convolutional Siamese network. Sustainability **12**(6), 2482 (2020). https://doi.org/10.3390/su12062482
11. Wei, Y., Wang, Z., Xu, M.: Road structure refined CNN for road extraction in aerial image. IEEE Geosci. Remote Sens. Lett. **14**(5), 709–713. 1027. https://doi.org/10.1109/LGRS.2017.2672734
12. Wang, S., Mu, X., Yang, D., et al.: Road extraction from remote sensing images using the inner convolution integrated encoder-decoder network and directional conditional random fields. Remote Sens. (2021).https://doi.org/10.3390/rs13030465
13. Alshehhi, R., Marpu, P.R., Woon, W.L., et al.: Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. Isprs J. Photogramm. Remote Sens. **130**(aug.), 139–149 (2017). https://doi.org/10.1016/j.isprsjprs.2017.05.002
14. Ganapathy, P., Skipper, J.A.: A novel ROC approach for performance evaluation of target detection algorithms. In: Conference on Automatic Target Recognition XVII. Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 207 Russ Engineering Center, 3640 Colonel Glenn Hwy, Dayton, OH 45435 (2007)
15. Alshaikhli, T., Liu, W., Maruyama, Y.: Simultaneous extraction of road and centerline from aerial images using a deep convolutional neural network. Int. J. Geo-Inf. (3) (2021). https://doi.org/10.3390/IJGI10030147
16. Pereg, D., Cohen, I., Vassiliou, A.A.: Sparse seismic deconvolution via recurrent neural network. J. Appl. Geophys. **175**, 103979 (2020). https://doi.org/10.1016/j.jappgeo.2020.103979



(a)Completeness (b)Correctness (c)F1

Fig. 4. Curve comparison for different methods.

5 Conclusion

Target detection in aerial images has been widely applied in many fields, including agriculture, forestry, electric power, land resources, urban planning, etc. In the acquisition process of aviation data, aircraft or UAV are constrained by the external environment, stability, wind resistance ability and clarity are limited, jitter phenomenon often occurs, camera Angle changes, etc. These uncertain factors will directly lead to the difficulty of road extraction.

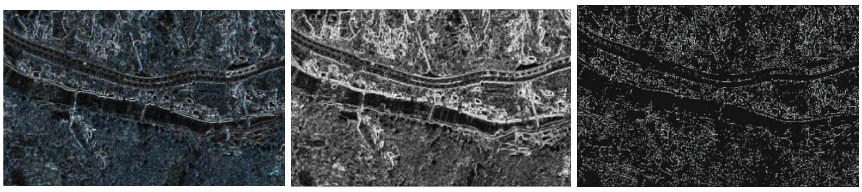
In this paper, a semi-automatic framework combining DCNN and SAE is studied to extract road information from aerial images. SAE model is used to learn the correlation between complex data, and the brief expression is found from the feature perspective. The decoder network samples the feature map extracted from the encoder network back to the input image of the same size, and finally the correct classification output is obtained by softmax classifier. Experimental results show that the proposed algorithm reduces the complexity of the model and improves the speed of calculation.

Acknowledgment. This work was supported in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01-03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474, and by Science and Technology Program in Xi'an city under Grant 21XJZZ0055.

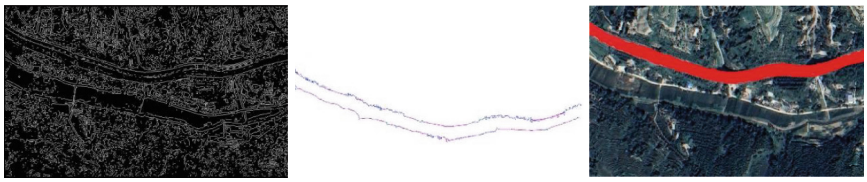
References

1. Eerapu, K.K., Lal, S., Narasimhadhan, A.V.: O-Seg-Net: robust encoder and decoder architecture for objects segmentation from aerial imagery data. IEEE Trans. Emerg. Top. Comput. Intell. **PP**(99), 1–12 (2021). https://doi.org/10.1109/TETCI.2020.3045485
2. Abdollahi, A., Pradhan, B.: Road extraction from open-source remote sensing dataset based on the modified deep convolutional autoencoders model. In: 43rd COSPAR Scientific Assembly Sydney, Australia, 28 January–04 February 2021. 2021
3. Tabibi, Z., Schwebel, D.C., Zolfaghari, H.: Road-crossing behavior in complex traffic situations: a comparison of children with and without ADHD. Child Psychiatry Hum. Dev. 1–8 (2021). https://doi.org/10.1007/s10578-021-01200-y

(a)Input (b)binary (c)Otsu



(d)Prewitt (e)Sobel (f)Valley



(g)Canny (h)Our method (i)Ground truth

Fig. 3. Comparison of different methods

Table 1. Objective Comparison

Methods	Completeness			Correctness			F1		
	aver	max	min	aver	max	min	aver	max	min
Ref. [4]	0.587	0.735	0.418	0.534	0.851	0.376	0.553	0.755	0.396
Ref. [6]	0.596	0.807	0.329	0.544	0.758	0.327	0.560	0.749	0.408
Ref. [8]	0.619	0.871	0.448	0.549	0.786	0.308	0.574	0.773	0.387
Our method	0.906	0.926	0.864	0.901	0.929	0.859	0.903	0.926	0.861

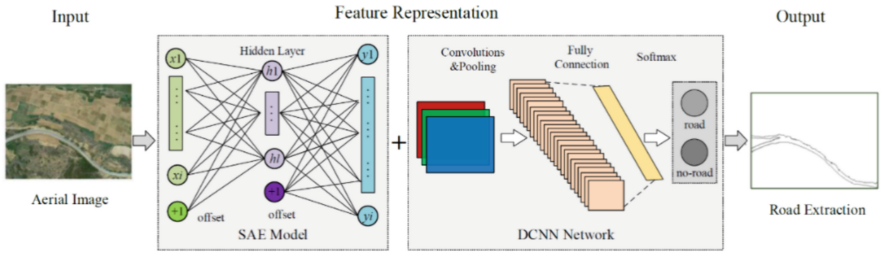


Fig. 2. Framework

4 Experimental Result and Analysis

4.1 Dataset Description

The experimental results of the above network framework are as follows. The dataset consists of two categories (urban roads and rural roads) with 900 images per category, where 400 images for training and 50 images for each group. For each image, the classification of ground truth is annotated by manual with the advice of experts carefully.

Through a large number of experiments, different initial values are selected, and the parameters with the highest performance in the cluster are selected to complete the network design. The evaluation system including Completeness, Correctness and F1 is used to test the road extraction performance. The formula can be denoted as in Eq. (4):

$$Com = \frac{TP}{TP + FN} \; Cor = \frac{TP}{TP + FP} \; F_1 = 2\frac{Com \times Cor}{Com{+}Cor} \tag{4}$$

where, *Com* means the completeness of matched with *GT* (ground truth) calculated by *TP* (truth positive) and *FN* (false negative), and *Cor* is correctness of matched with ground truth by *TP* and *FP* (false positive). F_1 is an overall that combines *Com* and *Cor*.

4.2 Result and Analysis

The proposed method is compared and to test robustness and flexibility of the related methods. The showing example from the testing dataset are shown in Fig. 3. In our testing images, the images were numbered as follows.

Figure 3 shows the different ways to achieve the road extraction, Table 1 gives the objective comparison of the results using Completeness, Correctness and F_1. Figure 4 is the line chart, which could exhibit the objective comparison more intuitively. In terms of Completeness, Correctness, and F1-score, the proposed method gives the best result in general.

The test can verify that the proposed method has some advantages compared to some existing methods in this field, and the results by our method is very close to ground truth, which are higher than the other methods. But for the Correctness, it is Performance is a bit poor which is caused by the almost indistinguishable gray level between the roads and the background in the bottom of the image.

3.2 Deep CNN

Methods based on deep learning have aroused widespread concerns, which establish a high-level semantic mapping relation by extracting the features. As a kind of feed-forward deep learning network, the Deep CNN is suitable for image feature extraction and recognition [15]. Usually, CNN architecture includes convolutional, mapping, pooling, fully connected, and output layers, that can be formed by stacking multiple underlying network structures.

First, feature extraction is performed in convolutional layer, and the formula can be denoted as in Eq. (3):

$$y_i = b_i + \sum\nolimits_i k_{ij} \ddot{A} x_i \tag{3}$$

where, y_i means the output image, x_i means the input image, $\otimes$ denotes convolution operator and k_{ij} is kernel function, finally, b_i is deviation value.

Second, the mapping layer employs a nonlinear activation function to obtain the feature map from the convolutional layer. The commonly used activation function is ReLU, sigmoid, tanh and softplus. Usually, the ReLU (Rectified Linear Units) function is employed as the activation function because the output will be zero, which could reduce the network and smooth the over-fitting problem.

Then, the pooling layer could avoid over-fitting phenomenon and maintain spatial invariance. And the full connection layer connects to all the previous layers including convolution layer or another full connection layer.

To train the network as a better performance, some operators, such as the local response normalization and dropout regularization method, are added to optimize results and speed up the training process. It randomly reduces the output of some neurons and reduces the neurons in the network that are no longer involved in the computation.

Finally, the classifier layer with full link is used to output in probabilistic form for each category. The most used loss function output in is the softmax function.

3.3 Framework

Therefore, a semi-supervised based deep learning method was proposed. The specific steps are as follows: Feature extraction part adopts the SAE model to study and find out the relationship between the optimal, get a concise expression, DCNN decoder of network from the encoder on the extraction of feature mapping samples back to the same size of the input image, and finally, at the end of the DCNN network using softmax classifier for the probability of road pixels in the final output.

Figure 2 shows the framework of our proposed method. The features learned by SAE, are applied to the convolution of a large number of training sets and test sets. The proposed DCNN network layers include one input, five conventional, two pooling, and one output. A max-pooling operation is performed between layer 1 and Layer 2. The average pooling layer follows the convolution of the five layers.

3.1 SAE Model

The performance of image classification is largely depended on the pros and cons of extracted features. SAE model is more suitable for unsupervised learning, which does not need a large number of tags during training massive aerial images. It can avoid the annotation of massive remote sensing images, and greatly improve the automation of the method. The unnecessary of annotation work can greatly improve the automation and efficiency of the algorithm [16].

The classic structure of SAE usually includes an input layer, in Fig. 1, a hidden layer, and an output layer, where +1 is the offset term.

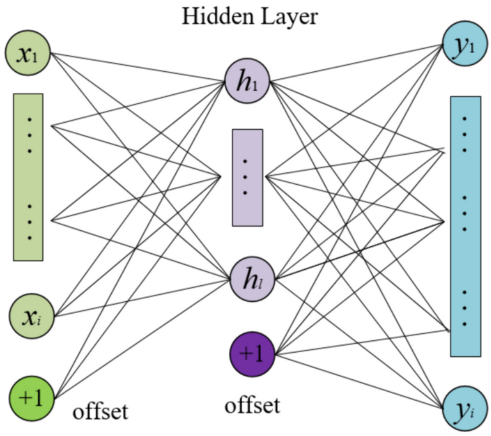


Fig. 1. SAE model

The loss function for neural network can be denoted as in Eq. (1):

$$J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}(x^i)-y^i\right\|^2\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^l)^2 \tag{1}$$

where, m is the amount of input samples, $(W,\ b)$ is the network parameter, n_l stands for the layers amount, s_l denotes the node amount in L layer, λ means the regularization and $h_{W,b}(x^i)$ means the output sample.

The SAE algorithm constrains the output of the hidden layer, so that the average could be high as 0. The loss function of SAE algorithm can be denoted as in Eq. (2):

$$J_{sparse}(W,b)=J(W,b)+\beta\sum_{j=1}^{s2}KL(\rho\|\hat{\rho})\tag{2}$$

where, ρ stands for the sparse parameter, $KL(\rho||\hat{\rho})$ measures the distributions.

Therefore, due to the aerial images have more complex backgrounds and targets. In this paper, a semi-automatic method combined Deep CNN with SAE (Sparse Auto-Encoder) is proposed to detect the road information from aerial image. First, the SAE model is carried out to learn the relationships and features of complex data and extract concise expressions from them automatically. Second, the encoder network completes the operation, and after the deep pooling and deconvolution operations, the intermediate features are extracted by the decoder network and sampled back to the input image of the same size on the map. Both convolution and pooling are implemented to reduce model complexity and boost distance calculation. Third, the final output is obtained by using the classifier, which is the probability distribution in the image representing the likelihood that the pixels in each region belong to the road and the non-road.

2 Related Work

In recent years, many methods have studied on road extraction from aerial image. Pradhan [8] proposed an automatic road extraction method by the neural network, which was superior to many methods of previous studies due to their ability to incorporate both multi-source information. Soni [9] presented a neural network to extract roads by a variety of texture parameters, and followed by the road vectorization stage. Experiments were carried out on different IKONOS and Quick Bird sample images to prove the road extraction capability of the proposed method.

Moreover, Nguyen [10] proposed a road extraction scheme based on feature learning, using convolutional neural network to capture the local structure of the road network. Due to the powerful learning ability of CNN, the road extraction method that we proposed can obtain high quality results. Wei [11] introduced a concise CNN for road extraction in aerial image. The paper proposed a new loss function which integrates the road geometry information into the cross-entropy loss. Experimental results showed that the model could perform well in accuracy, recall, F-score and accuracy.

Also, Wang [12] adopted a single patch architecture to extract roads from high-resolution images. Alshehhi [13] proposed a CNN network with integrated structure based on Alex-Net and VGG-net. Due to the large network structure, Alex-Net paid attention to the information of the large area. VGG networks focused on local details because of their small size. In this work, the training, verification and testing of the current popular deep learning models under different parameters have a good foundation for the identify and extraction of large geological and scientific data such as roads and buildings [14]. The accuracy of the road extraction is significantly improved.

3 Methodology

In our work, a semi-supervised based deep learning method was proposed, which combines Deep CNN (Convolutional Neural Network) with SAE (Sparse Auto-Encoder) to detect the road information from aerial image. In this part, the detail description of the concrete algorithms applied in our network is shown at first, and the overall framework and the algorithm execution process are elaborated on the follow.

Deep Neural Network Based on Sparse Auto-Encoder for Road Extraction

Sheng Liu, Shuxiao Chang, Ting Cao(✉), and Xinyue Li

Department of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi, China
caoting@xaut.edu.cn

Abstract. Road extraction from aerial image has realistic significance for GIS data updating. In view of the complexity challenging for acquiring road information, this paper proposes supervised model that combines Convolutional Neural Network (CNN) with Sparse Auto-Encoder (SAE) to cope with the road extraction task. First, the road features are extracted from the amount of non-annotated data using SAE model that aim to train the road features using CNN principle with implementing convolution and pooling to reduce model complexity. Second, the encoder network completes the operation, and after the deep pooling and deconvolution operations, the intermediate features are extracted by the decoder network and sampled back to the input image of the same size on the map. Third, the soft-max classifier categorizes images into roads and non-roads. Finally, the experiments verify that the proposed method outperforms the traditional methods and could achieve the satisfy result.

Keywords: Road extraction · aerial image · Deep learning · Convolutional Neural Network · Sparse Auto-encoder

1 Introduction

Road extraction from aerial images has vital usage in many applications including geographic information system, intelligent transportation system, environmental security and protection [1]. Various road extraction approaches can achieve road extraction successfully when the road exhibit obvious contrast respect with the non-road areas [2]. However, when the road with complex situation, such as road vehicles, buildings, tree occlusion cases, road extraction often appears discontinuous or gaps [3]. It is still challenging to deal with shadow or occlusion, geospatial information (urban, suburban or rural), and image scales, and obtain full and smooth road network automatically [4].

With the rapid development of deep learning in recent years [5], road extraction can be regarded as a classification task to distinguish aerial image into the road areas and the background areas [6, 7]. The state-of-the-art Convolutional Neural Network (CNN) is viewed as a successful deep learning model. CNN has advantages in hierarchical learning that makes it more efficient in feature extraction and image classification.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 120–127, 2024.
https://doi.org/10.1007/978-3-031-65126-7_12

12. Wang, H.: Research on underwater image enhancement based on improved MSRCR algorithm. **10**(06), 74–78+85 2020
13. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. Int. J. Comput. Vis. **1**(4), 321–331 (1988)
14. He, K., et al.: Guided image filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. ECCV 2010. LNCS, vol. 6311, pp. 1–14. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15549-9_1
15. Cui, Q.N., Tian, X.P., Wu, C.M.: Improved algorithm of haze removal based on guided filtering and dark channel prior. **45**(05), 85–290 (2018)
16. Lindeberg, T.: Feature detection with automatic scale selection. Int. J. Comput. Vis. **30**(2), 79–116 (1998)
17. Haralick, R.M., Linda, G.S.: Computer and Robot Vision, vol. 1. Addison-Wesley Longman Publishing Co., Inc., Boston (1992)
18. Marr, D., Hildreth, E.: Theory of edge detection. Proc. R. Soc. Lond. Seri. B. Biol. Sci. **207**(1167), 187–217 (1980)
19. LeCun, Y., et al.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
20. Malik, J., Perona, P.: Preattentive texture discrimination with early vision mechanisms. J. Opt. Soc. Am. A **7**(5), 923–932 (1990)
21. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Process. Lett. **9**(3), 81–84 (2002)

Click to select an image and select the one you want to operate on. Click on 'Identify' to remove fog and recognize traffic signs on the fog map. The image after defogging and recognition will be generated in the right display box.

5 Conclusion

This project mainly focuses on the recognition of traffic sign images in haze weather. Combining the basic theory of digital image processing with traditional Retinex vision theory, research and improvement on image defogging are carried out. A Retinex defogging algorithm with guidance filtering influence factors is designed, and the recognition algorithm for traffic signs is analyzed and implemented. Effective information extraction is carried out on the image. The classification of traffic images is carried out on CNN pre trained networks, and training and learning are conducted using the BelgiumTS traffic sign dataset, Finally, our Convolutional neural network can get 93.89% recognition accuracy, but in practical application, this number still needs to be improved.

Acknowledgment. This work was supported in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01-03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, 2022JQ-264, and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474, and by Science and Technology Program in Xi'an city under Grant 21XJZZ0055.

References

1. Improved Dark Channel Based License Plate Recognition Method and System Implementation in Foggy Environment by Wu Tianyuan, Chongqing University of Posts and Telecommunications, 2019
2. Feng, S.G.: Research on traffic sign image recognition algorithm based on haze weather (2019)
3. FusedGAN: Moving foggy image restoration beyond the limits by Ren et al. (2021)
4. Recurrent Squeeze-and-Excitation Context Aggregation Net for Single Image Dehazing by Liu et al. (2020)
5. Xue, B., Li, W., et al.: Review on feature extraction of traffic sign recognition **40**(06), 1024–1031 (2019)
6. Huang, F.: Parallelization implementation of the multi - scale Retinex image enhancement algorithm based on a many integrated core platform. Concurr. Comput. Pract. Exp. **32**(22) (2020)
7. Jobson, D.J., et al.: A multiscale Retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans. Image Process. **6**(7), 965–976 (1997)
8. McCann, M.T., et al.: The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. IEEE Trans. Inf. Theory **56**(1) (2010)
9. Gao, F., et al.: A novel multi-scale Retinex algorithm for image enhancement. In: Proceedings of the 6th International Conference on Image and Graphics (2011)
10. Methods and Applications of Image Retinex Problem by Yang Xue, Lanzhou University, 2020
11. Xiaofang, W., Dengjie, F., et al.: SRCR image defog algorithm based on multi-scale detail optimization. **37**(09), 92–97 (2020)

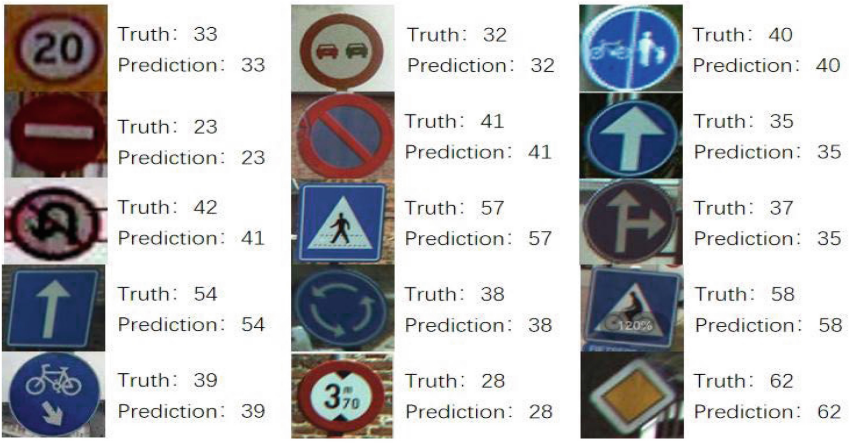


Fig. 8. Partial recognition results of improved defogging algorithm

4.3 Image Defogging and Traffic Sign Recognition System Based on Improved Retinex

The design of the improved Retinex based image defogging and traffic sign recognition system in this article is mainly divided into three parts: haze image selection, image display after defogging, and traffic sign box selection image display.

The Home screen of image processing consists of module selection and system menu, including four controls: button, panel, coordinate axis and text box. Each button has a corresponding callback function to switch the main interface to each module sub interface. The Home screen of GUI image processing system is shown in Fig. 9.

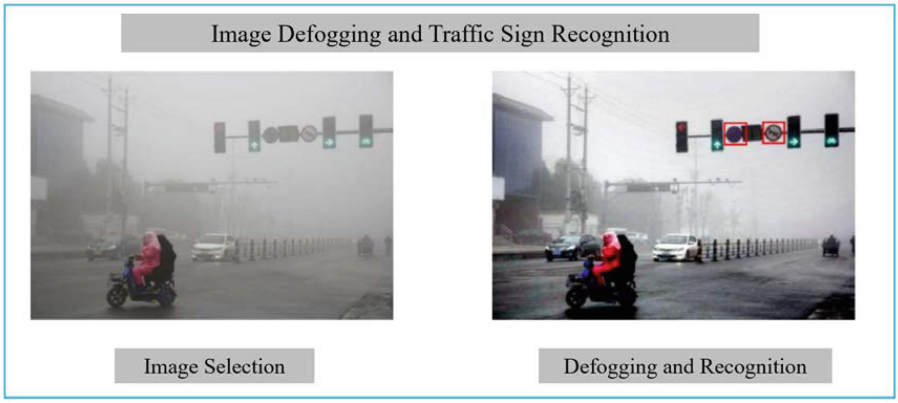


Fig. 9. Image Defogging and Traffic Sign Recognition System

Table 3. Experimental Results of the Second Group

algorithm	picture	Edge intensity factor	PSNR
SSR	pic (d)	41.0715	16.2509
MSR	pic (f)	42.0565	16.2701
Guided filtering	pic (h)	42.2539	16.7990
Weighted guided filtering algorithm	pic (j)	43.0994	16.8973

set graph prediction. In the figure, Input represents the true group identifier of the group of images to be predicted, while Prediction represents the predicted group identifier.

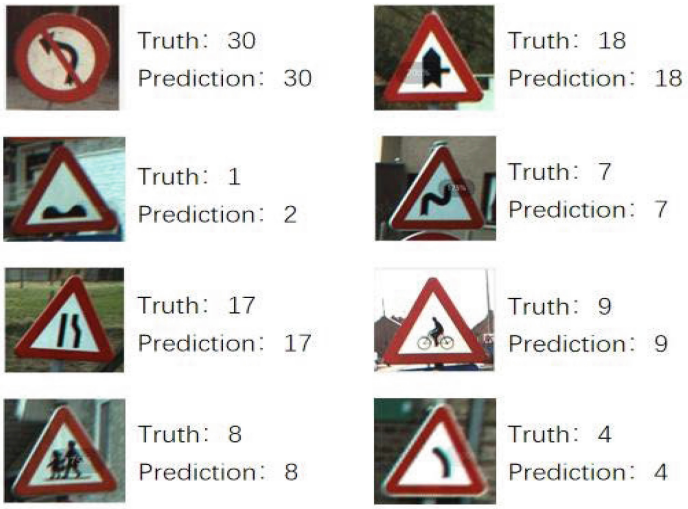


Fig. 7. Prediction Results of Test Set Part

It can be seen that there was an error in the recognition of the third image, but overall, the recognition rate can be maintained at a high level. After predicting all 2520 images in the test set, the accuracy of the prediction result is 92.73%.

Figures 8 show the prediction results of some traffic sign images processed by the improved Retinex defogging algorithm in this recognition algorithm, with two images showing recognition errors.

Figures 8 show the predicted results of some traffic sign images in the first set of experimental results of the defogging algorithm mentioned above. In this step, a total of 43 images from the first and second groups of the defogging algorithm results were used as predictive materials, with a total of 39 images predicting accurately, achieving an accuracy rate of 90.69%.

Fig. 5. Guiding Filter Processing Results

Fig. 6. Weighted Guided Filtering Processing Results

Table 2. Experimental Results of the First Group

algorithm	picture	Edge intensity factor	PSNR
SSR	pic (c)	82.8520	13.0193
MSR	pic (e)	83.5945	13.1975
Guided filtering	pic (g)	85.5671	13.9176
Weighted guided filtering algorithm	pic (i)	84.5239	13.8082

image noise. It can be seen that the experimental results of the algorithm combined with guided filtering contain more detailed information than the traditional Retinex algorithm. Therefore, it can be concluded that the improved Retinex algorithm can achieve edge information preservation to a certain extent.

4.2 Presentation and Analysis of Experimental Results on Traffic Sign Recognition

This part uses the Convolutional neural network pre training model designed in Sect. 3 to predict the test set, as shown in Fig. 7, the terminal output results of some kinds of test

Fig. 2. Original image

Fig. 3. SSR processing results

Fig. 4. MSR processing results

Therefore, two parameters, edge intensity factor and peak signal-to-noise ratio (PSNR), are selected as the comparison criteria, the images were divided into two groups for processing in the experiment. The average values of the experimental results of the two groups of images are shown in Tables 2 and 3 [20, 21]:

Among them, the edge intensity factor reflects the amount of edge information contained in the image. The larger the edge intensity factor, the clearer the image edges and the more edge information it contains; PSNR represents the ratio of signal to noise and is often used to evaluate noise and signal strength. A larger PSNR indicates less

Table 1. Network Architecture Diagram

Hierarchical network	feature maps	Convolutional kernel/pooling size	step
Conv2D	64 × 5	5 × 5	5
MaxPool2d	32 × 5	2 × 2	2
Conv2D	32 × 5 × 5	5 × 5	5
MaxPool2d	16 × 5 × 5	2 × 2	2
Fully connected layer	120	–	None
Fully connected layer	84	–	None
Fully connected layer	62	–	None

sets of images, including a total of 2520 images. The design and implementation of the training model are coded in the Python environment and PyTorch dependency package, which includes two convolutional layers, two maximum pooling layers, and three fully connected layers. A CNN network is implemented to achieve traffic image classification.

The structure of the designed network training model is shown in Table 1.

As shown in the above figure, after each convolutional layer is extracted, a pooling layer is added to reduce the dimensionality of feature information, thereby reducing computational complexity and accelerating network training speed.

After the training is completed, the model parameters are used to predict the test set, and the prediction accuracy for the test set can reach 92.73%. But for practical application requirements, this accuracy is not high. The CNN network model designed in this section is only a simple pre trained network, and its recognition performance still has room for improvement.

4 Experimental Results

4.1 Display and Analysis of Experimental Results of Defogging Algorithm

Here we use traditional single scale Retinex algorithm, multi-scale Retinex algorithm, and improved Retinex algorithm for experiments. In the traditional Retinex algorithm experiment, different Gaussian scales c are continuously adjusted to achieve better processing results. The final scale selection is: the scale in the single scale SSR algorithm is set to 15% of the image size; The multi-scale MSR algorithm has a mesoscale setting of 5% of the image size for small scales, 15% for medium scales, and 40% for large scales.

As shown in Fig. 2, the experimental example is shown in the original image. 4–2 shows the processing results of the single scale SSR algorithm, 4–3 shows the processing results of the multi-scale MSR algorithm, 4–4 shows the guided filtering processing results, and 4–5 shows the weighted guided filtering processing results (Figs. 3, 4, 5, 6):

It can be seen that both the traditional Retinex algorithm and the improved guided filtering algorithm can achieve good defogging results, achieving image enhancement results. However, the processing quality of edge information in each group of experimental results is difficult to compare with the naked eye.

1. Read in the guidance image I and the pending image P;
2. Calculate the mean and variance of I, the mean of the image P to be processed, and the product IP of I and P;
3. Calculate the linear correlation coefficient based on this $a_k = (IP - I_{mean}P_{mean})/(I_{Var} + \varepsilon)$; $b_k = P_{mean} - aI_{mean}$;
4. Calculate the mean of a_k and b_k;
5. Export filtering results: $q = a_{mean} * I + b_{mean}$;

This method uses guided filtering instead of Gaussian filtering to estimate illuminance images, ultimately resulting in an improved Retinex algorithm.

3.3 LOG Filtering Method

The edge detection algorithm of images requires both noise suppression and accurate positioning of edge information, and the LOG filtering method is an effective edge detection method. The LOG filter operator, also known as the Laplacian of Gaussian operator, and its corresponding operator, also known as the LOG operator, is the optimal filter for detecting image edge information based on image signal-to-noise ratio. This method comprehensively considers noise suppression and edge detection [16–18].

Perform the LOG operator on the test image to extract edge information, and the effect is shown in Fig. 1.

Fig. 1. Edge information extraction using LOG operator

Due to the extraction results being not suitable for observation, the pixel values of the resulting image were increased by a bias of 20 for observation. It can be seen that the LOG operator can effectively extract edge information from images.

3.4 Traffic Sign Image Recognition Based on CNN Network

This part will use Convolutional neural network (CNN) based on deep learning to implement a traffic sign image classification and recognition algorithm [19]. The training data is based on the BelgiumTS traffic sign dataset, which includes both training and testing sets. The training set contains 62 sets of images, each containing a certain number of logo images for training, with a total of 4575 images; The test set is also divided into 62

3.2 Improved Retinex Algorithm

The traditional Retinex algorithm uses Gaussian filtering for implementation, which has the effect of smoothing the image after processing. During the enhancement process of the image, there will be a loss of detail information, resulting in blurred information in the logo. So when calculating the illumination information of an image, we use guided filtering to estimate the illumination information of the image. Guided filtering is an edge preserving filter, and here we use guided filtering for illumination estimation [13, 14].

$$q_i = \sum_j W_{ij}(I) p_j \tag{1}$$

where p is the input image to be processed; I is the guiding image; q is the filtered output; W_{ij} is the filtering kernel, equivalent to $F(x, y)$, W_{ij} in the traditional Retinex algorithm is a function of guide image I. In actual calculations, we generally consider the output image q as the linear calculation result of guide image I. Assuming The output and input of the $W_{ij}(I)$ function are in a two-dimensional window Satisfy linear relationship within W_k:

$$q_i = a_k I_i + b_k, \forall i \in w_k \tag{2}$$

Among them, a_k and b_k is the constant term coefficient that needs to be calculated by us, and it is also the coefficient when the window center is located at k; w_k is the window; i and k are pixel indices.

As a local linear model, guided filtering is defined as the following Loss function in order to find linear correlation and minimize the difference between the output value of the fitting function and the true value p:

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left((a_k I_i + b_k - p_i)^2 + \varepsilon a_k^2 \right) \tag{3}$$

ω_k is right for a_k Correction compensation when the value is too large; The parameters about ε are used to adjust the blurriness of the image and the detection accuracy of edge information; εa_k^2 is used to suppress a_k value is too large. In terms of results, if the guide map does not contain edge information, the corresponding output mean filtering fuzzy result; If the guide map contains more edge information, the edge information will be reflected in the output image to achieve the preservation of edge information [15]. When calculating the coefficients of each window, a single pixel is usually described by multiple calculated linear functions. When calculating the output value of a single point, we take the mean of all calculated coefficients, and the final output result is as follows:

$$q_i = \frac{1}{|\omega_k|} \sum_{i \in \omega_k} (a_k I_i + b_k) = \overline{a}_i I_i + \overline{b}_i \tag{4}$$

Calculate the value of the linear coefficient from this. The algorithm flow of the guided filter is as follows:

Among the existing defogging algorithms, the retinex method has good performance and adaptability in image defogging. However, traditional retinex methods have drawbacks such as high computational complexity, difficulty in parameter selection, and limited effectiveness [2]. To address the problems of traditional retinex, we propose an improved retinex image clarity algorithm for image defogging and recognition of traffic signs in the image.

2 Related Work

Wang introduced traffic sign images into Convolutional neural network as training data to realize the classification function of traffic sign images. This method uses Convolutional neural network to study and classify the features of traffic signs, and can accurately recognize and classify different types of traffic signs. Li et al. proposed a method called FusedGAN to overcome the limitations of traditional defogging algorithms. This method combines the Generative adversarial network (GAN) and traditional image defogging technology, and restores images with complex haze by introducing multi-scale and multi-channel information fusion. FusedGAN can better remove the haze effect in the image and improve the image clarity and contrast [3]. Liu et al. proposed a single image defogging method based on the Recurrent Squeeze and Extraction Context Aggregation Network (R-SECA-Net). This method improves the quality and detail retention ability of image defogging by introducing attention mechanism and context aggregation. R-SECA-Net can adaptively adjust defogging processing, effectively reducing the problem of detail loss caused by haze [4]. Huang et al. proposed a traffic sign recognition algorithm based on CNN networks [5, 6]. The algorithm uses Convolutional neural network to extract and classify the features of traffic sign images, which can achieve high accuracy of traffic sign recognition.

3 Methodology

3.1 Traditional Retinex Algorithm

Among traditional Retinex image enhancement algorithms, the most common ones are single scale SSR algorithm, multi-scale MSR algorithm, etc., followed by iterative McCann algorithm and multi-scale Retinex algorithm with color restoration (MSRCR) [7–9]. The SSR algorithm needs to maintain a balance between contrast and image features, but the images to be processed vary in terms of shooting environment and imaging results. Therefore, the MSR algorithm is proposed based on the single scale algorithm. In order to compensate for the color deviation caused by interference such as haze and noise, a color restoration factor parameter C is added to the multi-scale MSR algorithm to adjust for the color deviation problem caused by the enhancement of local area contrast in the image. This corresponding algorithm is called the multi-scale Retinex algorithm with color restoration (MSRCR) [10–12].

Research on Traffic Sign Image Recognition Algorithms Under Complex Weather Conditions

Sheng Liu, Liming Qi, and Ting Cao(✉)

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
caoting@xaut.edu.cn

Abstract. In the transportation system, the influence of haze is more significant, such as license plate recognition, real-time monitoring, etc. The visibility of both people and equipment is greatly affected in foggy weather, leading to the emergence of foggy image processing. We analyzed the recognition requirements of traffic signs in foggy weather and conducted research on algorithms for removing fog from foggy images and extracting image edges. This topic mainly improved on the traditional Retinex algorithm, recognizing the loss of detail information in images under Gaussian filtering conditions. We applied guided filtering to the estimation of illumination images to achieve the preservation of image edge information. In terms of image recognition, the currently best performing LOG operator and Canny edge extraction algorithm were applied to achieve the extraction of detail information. Then, based on the background knowledge of Convolutional neural network, a small Convolutional neural network model is designed for training to realize the recognition and classification of traffic sign images. The experimental results show that the method proposed in this paper can achieve good functions in fog removal and traffic sign recognition.

Keywords: Haze · Edge detection · Retinex · Guided filtering · LOG operator · Convolutional neural network

1 Introduction

Traffic signs are the most important source for drivers to obtain road information during driving. As an important auxiliary facility in the road traffic system, traffic signs play an irreplaceable role. Haze inevitably reduces atmospheric visibility, and the accuracy and timeliness of driver information acquisition will be greatly negatively affected. Countless traffic accidents occur every year as a result. In addition, in foggy weather, the implementation of technologies such as intelligent monitoring, intelligent recognition, automatic navigation, and target tracking in outdoor environments has a significant negative impact. Therefore, in order to minimize the impact of haze weather on images, studying the implementation of traffic sign recognition algorithms under haze weather has extremely important theoretical value and practical significance [1].

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 109–119, 2024.
https://doi.org/10.1007/978-3-031-65126-7_11

4. Zhi-Liang, D., Yi-Qun, P., Jiantong, X., et al.: Application of reinforcement learning algorithm in operation optimization of air conditioning system . Build. Energy Conserv. **7**, 7 (2020)
5. Wang, X., Wu, J., Liu, C., Yang, H., Du, Y., Niu, W.: Fault time series prediction based on LSTM recurrent neural network. J. Beijing Univ. Aeronaut. Astronsut. **44**(4), 13 (2018)
6. Zhao, S., Dong, X.: Research on speech recognition based on improved LSTM deep neural networks. J. Zhengzhou Univ. Eng. Ed. **39**(5), 5 (2018). (in Chinese)
7. Ren Zhihui, X., Haoyu, F.S., et al.: Chinese lexical segmentation for sequence annotation based on LSTM network. Appl. Res. Comput. **34**(5), 5 (2017)
8. Rao, Q.: Research on Dimension Emotion Recognition based on Context. Jiangsu University (2017)
9. Lei, X.: Research on Short-time Vehicle Flow Prediction Model based on Integrated LSTM. Chongqing University of Posts and Telecommunications (2019)
10. Zhang, X.W., Li, Y.Y., Huang, S., et al.: Population situation prediction method of COVID-19 based on LSTM. CN111798991A (2020)

data vehicles in the training set. Although the relative error of a few predicted values is slightly larger, most of the results can meet the requirements.

This paper also introduces the mean absolute percentage error (MAPE), which indicates the accuracy of the model, and from the numerical values in the MAPE, how accurate the model is (Table 5).

Table 5. Lists the data.

Map	MAPE value
Working day	0.083704
Holidays and festivals	0.106437
Rainy day	0.126403
Traffic control day	0.107493

The average MAPE value of the calculated scene is 0.106, indicating that the average error is about 10.6% and the prediction accuracy is 89.4%, indicating that the model has a high precision forecast of short-term traffic flow.

5 Conclusion

As an important part of ITS, traffic flow forecasting system based on LSTM provides great help to people's travel and traffic management department's work. The system uses python language, uses Tensortflow + keras architecture to establish LSTM prediction model, and processes the acquired data set, which is divided into training set and test set. The prediction results of the training set are compared with the test set, and a good prediction accuracy is obtained. The model can be used to predict the future traffic flow, which provides help for people's travel and the work of relevant departments.

Acknowledgment. This work was supported in part by the Special Project of Technological Innovation and Guidance in Shaanxi Province under Grant 2022QFY01-03, in part by the Natural Science Foundation in Shaanxi Province under Grant 2022JQ-476, and in part by the Natural Science Foundation of Deduction Department in Shaanxi Province under Grant 2022JK0474, and by Science and Technology Program in Xi'an city under Grant 21XJZZ0055.

References

1. Wei, C.: Research on Taxi Shelter Design based on Service Design Concept. Xi 'an Polytechnic University (2017)
2. Liu, C.: Research on Urban Macro Travel Speed Prediction based on Classical Model and LSTM Model. Beijing Jiaotong University (2019)
3. Bonnin, R.: TensorFlow Machine Learning Project. Yao Pengpeng, trans. Beijing: People's Posts and Telecommunications Press (2017)

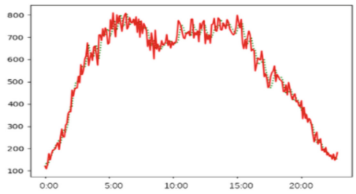


Fig. 2. Forecast of weekday traffic flow

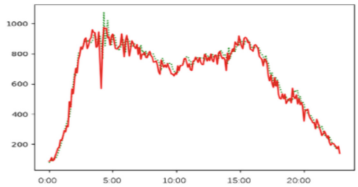


Fig. 3. Forecast of holiday traffic flow

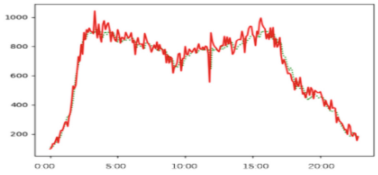


Fig. 4. Forecast of vehicle flow in rainy days

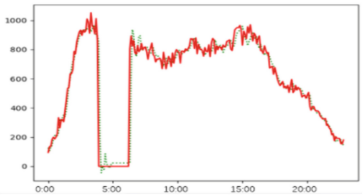


Fig. 5. Flow prediction under traffic control

which are brought into the model for training, and the prediction results are obtained. Use the plotting function to represent the prediction results. Among them, the real value is set as the red solid line, and the prediction result is the green dotted line. It can be seen from the figure that the short-term vehicle flow prediction model based on LSTM has a better vehicle flow prediction result for working days.

In general, the relative error of the sequence value of traffic flow on the predicted date can be obtained from the figure. To a certain extent, the main reason for the error in the analysis is that the model only considers the time characteristics of vehicle flow while other influencing factors are not taken into account, and there are not enough

3.2 Neural Network Training Steps for Long and Short Time Memory

Based on the above error propagation formula, the training steps of LSTM neural network are as follows:

Firstly, the output value of each neuron is calculated by the forward calculation formula, and the output value of the whole network is calculated according to the output value of each neuron;

The total output is compared with the actual value, the total error value is obtained, and the deviation of each neuron is calculated by the back propagation algorithm; The gradient descent algorithm is used to calculate and update each weight gradient according to the corresponding error value; After the new network weight is obtained, the forward calculation formula is continued to calculate the output of each neuron according to the new input data, and the final output of the network is obtained and then compared with the actual value.

When a certain error accuracy is reached, the network parameter is saved. At this time, when the network training is completed, if the error accuracy is not reached, the iteration is carried out continuously. Until the network output reaches a certain error accuracy.

4 Experimental Results and Analysis

4.1 Software Running Environment and Hardware Configuration

In terms of hardware environment, the Windows OS version is win11, the operating system is 64-bit, the processor model is Intel i5 8300 h, the display adapter is NVIDIA GTX1050ti, and the memory is 16 GB. In terms of software environment, the programming language is python, the python version is 3.6.8, and pycharm is selected as the integrated development environment.

4.2 Operation Results and Analysis

The built training set data was fed into the prediction model, and four different datasets were built using real-time data collected from individual detectors in the highway system across California's metropolitan areas between April 8 and May 10, 2018, classified by four different conditions: weekdays, holidays, rainy days, and traffic control days. Each part of the data is 4-day traffic flow data, and the sample set is built with the number of vehicles passing through high-speed cameras every 5 min. The vehicle flow prediction model of a short-duration memory network is constructed and the trained model is saved. The model input is obtained by the sequential sampling method described above, and then the vehicle flow of the same day is predicted on a rolling basis.

Here, plt.plot() function in Matplotlib is used to plot the prediction results and the real values, and 4 graphs of the traffic flow prediction results under different scenarios are obtained. Figures 2, 3, 4 and 5.

As shown in the figure above, the traffic flow of the following day is predicted based on the traffic flow data of the previous three days, and the total number of samples is 1152 each time. The 864 samples of the first three quarters are used as training sets,

The LSTM neural network uses two gates to control the cell state c, one is the amnesia gate, whose function is to judge how many cell states exist from the last time c_{t-1} to the time c_t, and the other is the input gate. Its function is to determine how much timely input x_t to the cell state c_t. Another gate is the output gate [9], whose function is to determine how much cell state c_t is output to the LSTM's current output value (Fig. 1).

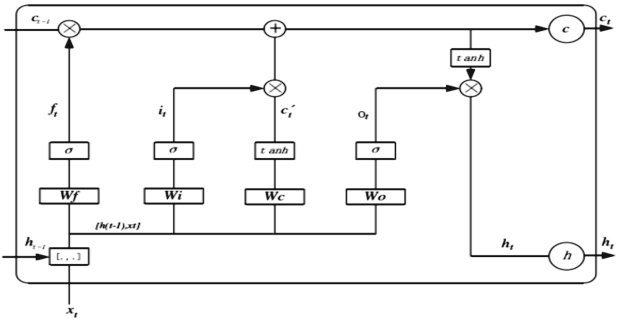


Fig. 1. Internal calculation flow diagram of a short-duration memory neural network unit

In the LSTM network, forward propagation and computation of information are also carried out through neuron transmission, and the network forward computation can be expressed by six formulas [10]. The first is the forgetting door and the calculation of the forgetting door is:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{3}$$

The input gate is calculated as follows:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{4}$$

The unit state describing the current input is calculated based on the output of the previous moment and the input of the current moment, and its calculation expression is as follows:

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{5}$$

$$\widetilde{c_t} = tanh(W_c \times [h_{t-1}, x_t] + b_c) \tag{6}$$

The number of gates that control the long-term memory acting on the current instantaneous output is output gate o_t, and its calculation formula is (7).

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \tag{7}$$

The output result of the cell is determined by the output gate o_t and the cell state c_t, expressed as follows:

$$h_t = o_t * tanh(c_t) \tag{8}$$

The above formula 8 is the forward calculation expression of the whole long and short time memory network.

$$X_{scaled} = X_{std} * (max - min) + min \tag{2}$$

2.2 Model Output

In this paper, Tensorflow + keras architecture is used to construct the prediction model of long and short term memory network and train the neural network model with data set. As an open library [3], Tensorflow is a symbolic mathematical system based on data stream programming, which has been widely used in the research of machine learning, deep neural networks and other fields. Tensorflow is composed of multi-level institutions, which can use GPU and TPU for numerical calculation, and supports C and Python, which is completed in python language in Pycharm. In the process of using Tensorflow to build a neural network model, the environment should first be built. The version selected here is Tensorflow-GPU-2.6.0, and suitable CUDA should be installed. Secondly, various necessary packages should be imported into Pycharm. Such as keras, matplotlib, MKL sklearn etc., after setting various parameters, in this paper, the model set to 2 layer LSTM model. The process of constructing LSTM neural network using Tensorflow library is as follows: First, build a Sequential model with Sequential() to add layers. When building the neural network model, it is necessary to set the parameters of the model in the program, which has been introduced before. The number of hidden neurons in the first layer LSTM network is 50 and the output dimension is 50. Return_sequences is set to True and only the output of the last state is returned. The output dimension of the second layer is 100, the output dimension of the Dense layer is 1, and the activation function is liner.

After the traffic flow prediction model is built, it is necessary to train the neural network model with the sample set. In this paper, the loss function loss is defined as the mean square error function. RMSprop is used in the optimizer and batch gradient descent algorithm is adopted. According to the above, the training batch size is 8, which means that 8 data are extracted from the training set at a time for model training. During the experiment, the data of the training set was continuously converted into the prediction model, and the prediction results were compared with the real data of the verification set [4]. After the deviation was obtained, the backpropagation algorithm was used to optimize the training network, and the model was continuously trained. In this paper, 300 iterations of training were set as the end condition.

3 Related Technology

3.1 The Internal Structure of Long - Term Memory Network and the Calculation Method of Data

Long term memory neural network has completely overcome the shortcomings of ordinary RNNS [5], and is the most widely used RNN at present, which is widely used in many fields such as speech and picture recognition [6], natural language processing [7], emotion recognition [8] and so on. In the LSTM neural network, in addition to the short-term input signal sensitive state h, the cell state c is added and used to store the long-term state.

data of the first three days are constructed and trained as a sample set and retained as a vehicle flow prediction set of the following day to test the accuracy of the model.

In this paper, vehicle data at n time intervals before the road surface is used to estimate vehicle data at a subsequent time unit. Therefore, the sample set is constructed by inputting data at n time from past time into the network, where X represents the input network arrangement, the data at n + 1 time becomes the network output, and Y is taken as the network output arrangement. Neural networks generally associate input and output with information. After training, the network will form another input network and obtain a new output by direct association with specific data. Therefore, network input X and network output Y jointly construct the sample set of the experiment, and the data set is obtained in the form of rolling forward sampling, extracting n + 1 pieces of information at a time. Previously, n pieces of information were input X, and rolling prediction was also made in the prediction to build a larger number of samples. Therefore, in order to process the experimental data sequence into a short-duration memory network which can adapt to the data arrangement, it is necessary to use functions to complete the above requirements. There are a total of 3 * 288 = 864 experimental data in the three days from 4–8 to 4–10. When m + 1 experimental data is set at one time, the sample number constructed when m + 1 experimental data is used as the input of the experiment and m + 1 experimental data is used as the output of the experiment is (864-m) item. A data type in the constructed training sample set is shown in Table 4 (m = 11).

Table 4. Data after sequence transformation

X
……
130,155,125,124,111,117,91,119,79,112,81
155,125,124,111,117,91,119,79,112,81,88
125,124,111,117,91,119,79,112,81,88,71
124,111,117,91,119,79,112,81,88,71,78
111,117,91,119,79,112,81,88,71,78,118
117,91,119,79,112,81,88,71,78,118,82
91,119,79,112,81,88,71,78,118,82,71
119,79,112,81,88,71,78,118,82,71,83
……

Since the sigmoid function is used in the hidden layer of the experimental network, in order to achieve the speed of network convergence and prevent the problem of neuron saturation, it is generally required to normalize the training information. Here, MinMaxScaler() function in numpy library is selected to normalize the data.

$$X_{std} = \frac{X - X.\min(axis = 0)}{X.\max(axis = 0) - X.\min(axis = 0)} \tag{1}$$

The collected data includes six data types: Local Date, TIME, ID, CXDM, HPZL, and SYXZ. Table 2 describes the meanings of each data type.

Table 2. Original fields

fieldName	Local Date	TIME	ID	CXDM	HPZL	SYXZ
Meaning	Date	Record timestamp	Vehicle number	Vehicle code	Type of license plate	Nature of vehicle use

Based on the research content, part of the data is selected for extraction, and the data used are two types: Local TIME (date) and TIME (record time).

Based on the past traffic flow of a specific section of the road, the vehicle data of the next moment can be predicted, so that it is possible to train the neural network by constructing a data set based on the past traffic flow value. Before constructing the data set, the traffic flow value of the intersection should be extracted successively in time period. Sort the traffic according to the time sequence, and then use the resample() function in Pandas to summarize the time data in a period of 5 min when the traffic passes through the intersection. The summary results are shown in Table 3.

Table 3. Extracted data

Time	Volume
……	……
2018/4/8 0:05	130
2018/4/8 0:10	155
2018/4/8 0:15	125
2018/4/8 0:20	124
2018/4/8 0:25	111
2018/4/8 0:30	117
2018/4/8 0:35	91
2018/4/8 0:40	119
……	……

After pre-processing and time series value extraction, the traffic data passed by the intersection has been summarized into a five-minute time span of traffic flow data and stored in the document. All data starts at 2018-04-03 00:00:00 and ends at 2018-05-31 23:55:00. The data span is 5 min.

In this paper, the vehicle flow data of 30 days from midnight of April 3 to evening of May 10 are divided into four conditions according to the actual situation: working day, holiday, rainy day and traffic control. According to the situation, the vehicle flow

For the problem of traffic flow prediction, domestic and foreign scholars have conducted a lot of research, and built three research models, which are support vector machine model (SVM), deep learning model and neural network model (NN). Because of its own characteristics, SVM has been widely used in short-term traffic flow prediction, solving the problem of too small data and other problems. With the continuous development of deep learning, deep learning has also been studied in traffic flow prediction to a certain extent. The research on deep learning in traffic flow prediction is just at the beginning stage. According to the existing results, deep learning has high accuracy in predicting the results of specific data, but it also has certain limitations, and there are certain problems in the aspects of large calculation amount and long consumption time. Over the years, scholars have developed a variety of neural network models. Compared to previous machine learning, neural networks have more hidden units, using methods to abstract objects at a deeper level and extract hidden features that the data cannot see. Neural network also has the characteristics of strong adaptability.

On the whole, the development direction of short-term traffic prediction has gradually transited from linear model to nonlinear model, and from non-intelligent direction to intelligent prediction direction [2]. Based on the analysis of estimation accuracy, training model difficulty and reliability, the neural network model is usually better than the traditional model in the case of the same data, and is more suitable for short-term traffic flow prediction.

This paper will use the previous traffic data and build a model to predict the short-term traffic flow and analyze the results, so as to provide a helping hand to promote the smooth layout of ITS, reduce the current increasingly congested traffic situation, and ensure the safe and convenient travel of the people and the improvement of road travel experience.

2 Short Time Traffic Flow Forecasting Model

2.1 Model Input

The information used here is based on data from the Traffic Flow Prediction Project database collected in real time from individual detectors in the highway system across California's metropolitan areas. The time span is 1 month, 38 days from April 3 to May 10, 2018 (Table 1).

Table 1. Original data examples

Local Date	TIME	ID	CXDM	HPZL	SYXZ
2018/4/8	00:03:40	13631	K33	02	A
2018/4/28	14:38:29	41256	K31	02	A
2018/5/9	23:19:03	67293	K33	03	D
……	……	……	……	……	……

Design and Implementation of Traffic Flow Prediction Model Based on Short and Long Time Memory Network

Sheng Liu, Xinyue Li, Ting Cao(✉), and Shuxiao Chang

School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
caoting@xaut.edu.cn

Abstract. Due to the randomness, fuzziness, time variability and uncertainty of traffic flow, it is difficult for traditional forecasting models based on time series or artificial neural networks to accurately reflect the actual traffic situation, etc. This paper takes the demand for short-term traffic flow forecasting of urban rail transit as the research object, and analyzes the implementation methods suitable for short-term traffic flow forecasting. LSTM neural network was used to construct the model for simulation experiment analysis. The results of data analysis show that the LSTM neural network model obtains the minimum average absolute percentage error MAPE value of 10.6% and the highest average accuracy of 89.4%, which has a good prediction effect and can improve the prediction work of short-term traffic flow.

Keywords: Neural network · Integrated learning · LSTM · Short time traffic flow forecast

1 Introduction

In recent years, with the rapid development of our economy, the number of motor vehicles and non-motor vehicles in urban areas is also increasing. The overall number of motor vehicles has been increasing, and the problem of the sharp increase in urban traffic pressure has also followed. In order to solve many problems such as the worsening of urban traffic congestion, in the context of the continuous development of modern social science and technology, the Internet of Things and other emerging technologies have made people put forward more advanced ways to improve traffic conditions, and gradually formed the concept of intelligent transportation system (ITS).

ITS is an efficient management system, which relies on road engineering to reduce traffic congestion and natural pollution and ensure smooth traffic operation. Traffic flow prediction is a very key part of ITS [1], which can quickly help the system to achieve timely, dynamic, accurate and reliable quantitative prediction of vehicle data and future road traffic flow conditions.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 99–108, 2024.
https://doi.org/10.1007/978-3-031-65126-7_10

Transportation Networks

6. Deckersbach, T., Wilhelm, S., Keuthen, N.J., Baer, L., Jenike, M.A.: Cognitive-behavior therapy for self-injurious skin picking: a case series. Behav. Modif. **26**(3), 361–377 (2002)
7. Odlaug, B.L., Grant, J.E.: Clinical characteristics and medical complications of pathologic skin picking. Gen. Hosp. Psychiatry **30**(1), 61–66 (2008)
8. Tucker, B.T., Woods, D.W., Flessner, C.A., Franklin, S.A., Franklin, M.E.: The skin picking impact project: phenomenology, interference, and treatment utilization of pathological skin picking in a population-based sample. J. Anxiety Disord. **25**(1), 88–95 (2011)
9. Prochwicz, K., Antosz-Rekucka, R., Kałużna-Wielobób, A., Sznajder, D., Kłosowska, J.: Negative affectivity moderates the relationship between attentional control and focused skin picking. Int. J. Environ. Res. Public Health **19**(11), 6636 (2022)
10. Brämer, G.R.: International statistical classification of diseases and related health problems. Tenth revision. World Health Stat. Q. **41**(1), 32–36 (1988)
11. Guha, M.: Diagnostic and statistical manual of mental disorders: DSM-5. Ref. Rev. **28**(3), 36–37 (2014)
12. Cohen, L.J., et al.: Clinical profile, comorbidity, and treatment history in 123 hair pullers: a survey study. J. Clin. Psychiatry **56**(7), 319–326 (1995)
13. Woods, D.W., et al.: The trichotillomania impact project (tip): exploring phenomenology, functional impairment, and treatment utilization. J. Clin. Psychiatry **67**(12), 1877 (2006)
14. Higashi, S., Goto, D., Okada, S., Shiozawa, N., Makikawa, M.: Development of wearable EMG measurement system on forearm for wrist gestures discrimination. In: 2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech), pp. 250–251. IEEE (2019)
15. Turlapaty, A.C., Gokaraju, B.: Feature analysis for classification of physical actions using surface EMG data. IEEE Sens. J. **19**(24), 12196–12204 (2019)
16. Fajardo, J.M., Gomez, O., Prieto, F.: EMG hand gesture classification using handcrafted and deep features. Biomed. Signal Process. Control **63**, 102210 (2021)
17. Campanini, I., Disselhorst-Klug, C., Rymer, W.Z., Merletti, R.: Surface EMG in clinical assessment and neurorehabilitation: barriers limiting its use. Front. Neurol. **11**, 556522 (2020)

OCD presents a complex and heterogeneous mental health challenge, requiring in-depth investigation to comprehend its underlying mechanisms fully. This paper ventures into the unexplored territory of OCD and its potential link to human skin textures, with a specific focus on Excoriation Disorder, commonly known as chronic skin-picking. By conducting a comprehensive analysis of existing literature, this study seeks to elucidate the cognitive aspects, memory impairments, and potential neurobiological factors contributing to this unique association.

5 Conclusion

In summary, Obsessive-Compulsive Disorder (OCD) is a complex mental health condition challenging our understanding. This study explores the potential link between OCD and human skin textures, focusing on Excoriation Disorder (chronic skin-picking) through comprehensive literature analysis. Examining cognitive, memory, and neurobiological factors, it also considers human-computer interaction (HCI) in analysis and treatment, emphasizing skin texture aspects. Two avenues for understanding OCD through skin texture emerge: repetitive movements due to memory issues, resulting in enlarged objects with skin texture imprints, and identifying OCD patterns through distinctive skin marks from compulsive skin peeling. Inspired by Exposure and Response Prevention therapy, magnifying skin texture details simulates ERP, countering perfectionism. This innovative approach provides insights into OCD complexities, underlining skin texture's role in understanding and treating the disorder. By integrating cognitive and neurobiological aspects, this study advances our understanding of the intricate OCD-skin texture relationship, aiding OCD research and interventions for a comprehensive perspective on this condition.

Acknowledgment. This work is supported by the Shenzhen Science and Technology Innovation Commission (Stabilisation Support Programme).

References

1. Shusta, S.R.: Successful treatment of refractory obsessive-compulsive disorder. Am. J. Psychother. **53**(3), 377–391 (1999)
2. Savage, C.R., Baer, L., Keuthen, N.J., Brown, H.D., Rauch, S.L., Jenike, M.A.: Organizational strategies mediate nonverbal memory impairment in obsessive-compulsive disorder. Biol. Psychiat. **45**(7), 905–916 (1999)
3. Benzina, N., Mallet, L., Burguière, E., N'diaye, K., Pelissolo, A.: Cognitive dysfunction in obsessive-compulsive disorder. Curr. Psychiatry Rep. **18**, 1–11 (2016)
4. Law, C., Boisseau, C.L.: Exposure and response prevention in the treatment of obsessive-compulsive disorder: current perspectives. Psychol. Res. Behav. Manage. **12**, 1167–1174 (2019)
5. Grant, J.E., Chamberlain, S.R.: Trichotillomania and skin-picking disorder: an update. Focus **19**(4), 405–412 (2021)

precision score, recall score, f1 score, and accuracy. Compared results are shown in Fig. 3.

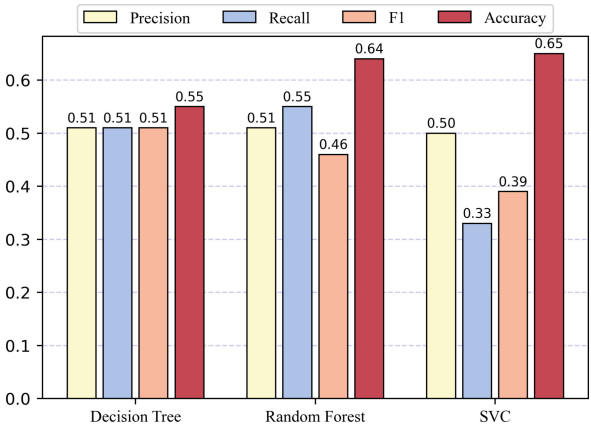


Fig. 3. Compared results of EMG classification on the WESAD dataset.

We can see that the SVC achieves the best accuracy which is 0.65 and the Decision Tree model achieves the best F1 score which is 0.51. We can also find that although the accuracy of SVC is high, its F1 is not high, because the number of samples in the two categories of the data set is unbalanced, and SVC is more likely to classify the samples to the one with a larger number.

4 Discussion

The findings carry important implications for OCD research and intervention. Understanding the connection between OCD and skin texture enhances comprehension of this intricate disorder. The immersive ERP approach offers a promising therapeutic avenue for directly confronting obsessions and compulsions.

Despite the study's strengths, certain limitations must be acknowledged. Sample size and specific OCD subtypes may influence the generalizability of the findings. Additionally, the accuracy and limitations of wearable devices can impact data quality.

In conclusion, this academic discourse explores the intricate relationship between OCD and human skin textures. Utilizing wearable devices and an immersive ERP approach, valuable insights into obsessive-compulsive traits through skin texture analysis are gained. The combination of EMG signals and skin features enables a holistic exploration of OCD's underlying mechanisms. This research contributes to the field by advancing our understanding of OCD and proposing potential skin texture-based interventions in OCD assessment and treatment.

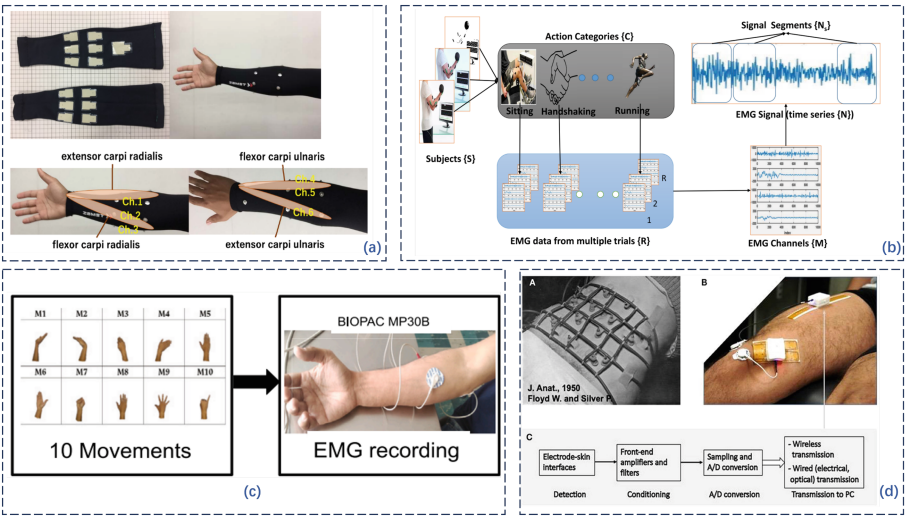


Fig. 2. Existing schematic diagram of AIoT Framework for EMG-based OCD detection.

In summary, the proposed framework aims to collect the EMG signal, then capture the correlation between OCD and EMG, and finally complete OCD detection. We also review some existing schematic diagrams of the AIoT Framework for EMG-based OCD detection as Fig. 2.

Figure 2 (a) shows a wearable EMG measurement system on the forearm for wrist gesture discriminations, in which twelve electrodes are utilized to measure EMG from six locations [14]. Figure 2 (b) shows surface EMG Data collection for the classification of physical actions, which observations consist of C categories with R trials [15]. Figure 2 (c) shows arm EMG signal collection for hand gesture classification, which contains 10 gesture classes, and each gesture is repeated 100 times by the participants during collection [16]. Figure 2 (d) illustrates EMG detection with wearable sensors for abdominal muscle monitorings in 1950 as shown in (A) and the modern system for knee EMG detection in (B), where (C) illustrates the system [17].

3.4 EMG Classification Case

To better understand the EMG signal, we provide a classification case with the Wearable Stress and Affect Detection (Wesad) dataset. This dataset provides EMG signals with four different emotion states: neutral, stressed, amused, and meditated. We first divided these states into neutral (neutral and neutral) as label 0 and active (stressed and amused) as label 1. Then we downsample the signal to 256 Hz and use a sliding window approach to obtain 10-second segments. Finally, we use Decision Tree, Random Forest, and Support Vector Classifier to make a classification. To evaluate the performance of models, we compare the

paving the way for the development of targeted interventions for affected individuals.

3.3 Proposed AIoT Framework for EMG-Based OCD Detection

We propose an AIoT for EMG-based OCD detection, which contains three main modules: signal collection, signal processing, and OCD detection. The proposed framework is shown in Fig. 1.

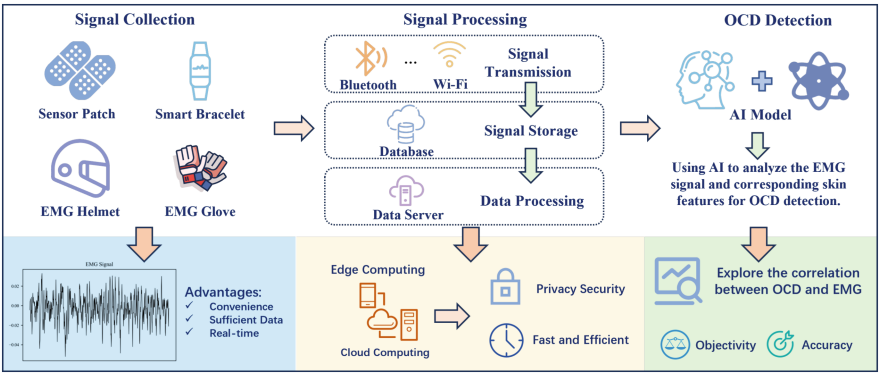


Fig. 1. AIoT Framework for EMG-based OCD detection. The proposed AIoT Framework for EMG-based OCD detection can be divided into three main parts: signal collection, signal processing, and OCD detection.

First, wearable devices are widely used for EMG signal collection, such as skin sensor patches, smart bracelets, EMG helmets, and EMG gloves. These devices can collect sufficient EMG data in a convenient way, and provide a way to collect data in real time. These wearable devices can also be personalized according to the needs of the user and the collected signal to meet the requirements of different users. Then, IoT techniques are used for EMG signal processing. Specifically, EMG signals are transferred with Bluetooth or Wi-Fi from wearable devices to the database for signal storage, and data servers are used for data processing with edge computing and cloud computing. In this way, it can provide fast and efficient signal storage and processing with privacy security. In addition, the database can store long-term signal data to achieve continuous recording, which can achieve better correlation mining. Finally, AI models such as CNNs and RNNs are used for analysis. Using AI models to analyze the EMG signal and corresponding skin features can help explore the correlation between OCD and EMG in an objective and accurate way. In addition, other AI technologies, such as big data, can also provide effective assistance to explore the correlation between OCD and EMG. With the correlation explored, a classification model can be constructed for OCD detection.

utilization of wearable technology enhances understanding of skin-picking disorder's complexities and its potential connections to OCD, potentially leading to more effective therapeutic approaches for managing these conditions.

3.2 Method for Trichotillomania

The research methodology revolves around data analysis using wearable devices, specifically focusing on EMG signals and corresponding skin features. This approach aims to explore the correlation between OCD and EMG data while integrating physiological and skin texture information to gain deeper insights into the association between OCD symptoms and changes in skin texture.

The study intends to utilize the EMG signals to investigate individuals with Trichotillomania, analyzing the muscle activity involved in hair-pulling. By capturing EMG signals and corresponding skin features, the research seeks to understand the patterns and frequency of hair-pulling episodes, as well as the physical consequences, such as hair loss and potential skin injuries. The analysis of this data may reveal connections between EMG signals, skin texture changes, and the severity of Trichotillomania symptoms, offering a comprehensive exploration of the disorder's impact on the skin.

Additionally, the wearable devices' data analysis will be extended to explore hair-pulling patterns across different body areas, such as the scalp, eyebrows, eyelashes, and pubic area, as mentioned in the previous. This investigation aims to identify any differences in the intensity or frequency of hair pulling in various regions, contributing to a more nuanced understanding of Trichotillomania and informing targeted interventions.

Furthermore, the duration of hair-pulling episodes will be assessed through the EMG signals, enabling researchers to understand the persistence and intensity of these behaviors. The study seeks to correlate the duration of hair-pulling episodes with the severity of Trichotillomania symptoms and potential skin damage, facilitating the development of interventions to interrupt hair-pulling behavior and promote healthier coping strategies.

The research methodology also aims to explore correlations between OCD symptoms and EMG data related to skin-picking or hair-pulling behaviors. By capturing EMG signals during episodes of skin picking and analyzing corresponding skin texture features, the study intends to investigate potential associations between the intensity of skin-picking behaviors, the severity of OCD symptoms, and observable skin changes. This integrated approach offers valuable insights into the relationship between OCD and Body-Focused Repetitive Behaviors (BFRBs) like skin picking, contributing to a deeper understanding of their underlying mechanisms and potential links between the disorders.

The research methodology, centered on wearable devices and data analysis of EMG signals and skin features, provides a unique opportunity to comprehensively explore the association between OCD symptoms and changes in skin texture in BFRBs like Trichotillomania and skin picking. The integration of physiological and skin texture information enhances our understanding of these disorders and may lead to advancements in both OCD and BFRB research,

3 Proposed Method

The research methodology is centered on data analysis using wearable devices. Electromyography (EMG) signals and corresponding skin features are captured to explore the correlation between OCD and EMG data. Integrating physiological and skin texture information aims to provide deeper insights into the association between OCD symptoms and skin texture changes.

3.1 Method for Skin-Picking Disorder

The research methodology revolves around data analysis using wearable devices, with a primary focus on capturing EMG signals and corresponding skin features. The objective is to explore the correlation between OCD and EMG data while integrating physiological and skin texture information to gain deeper insights into the association between OCD symptoms and skin texture changes. Through the application of wearable devices, the study aims to investigate individuals with skin-picking disorder, analyzing the EMG signals during episodes of skin-picking and correlating them with the corresponding skin texture changes. This analysis seeks to uncover patterns in EMG activity associated with skin-picking behaviors, contributing to the identification of potential markers of the disorder.

Moreover, the research methodology seeks to identify specific characteristics of OCD through the analysis of skin texture changes in conjunction with EMG data. By integrating physiological and skin texture information, the study endeavors to explore whether certain patterns in skin texture changes are associated with OCD symptoms, providing a deeper understanding of the underlying mechanisms.

Additionally, the research methodology aims to assess the psychological consequences of skin scratching by analyzing EMG data and corresponding skin features. The study intends to explore the correlation between the severity of skin-picking behaviors and the level of distress experienced by individuals, shedding light on the psychological impact of skin-picking and its potential role in the maintenance of the disorder.

The research methodology can contribute to the classification of skin picking as a Body-Focused Repetitive Behavior (BFRB) by capturing EMG signals during recurrent and habitual actions directed at the body. By establishing a link between the repetitive behaviors observed in skin picking and the broader category of BFRBs, the study provides a foundation for understanding the relationship between different BFRBs and contributes to a comprehensive understanding of these conditions.

In conclusion, the research methodology focuses on data analysis using wearable devices, particularly EMG signals and skin features, to comprehensively explore the association between OCD symptoms and skin texture changes in skin picking disorder. By integrating physiological and skin texture information, the study aims to advance knowledge in both OCD and BFRB research, informing the development of targeted interventions for affected individuals. The

able to control himself to think about this phenomenon over and over again; some obsessive-compulsive disorder sufferers cannot accept the imperfections on their skin, such as scars or pimple marks and thus repeatedly think about them and touch them, which not only affects their daily life, but also causes more damages to their skin itself in severe cases. Individuals with OCD demonstrate slowed performance on neuropsychological tests due to excessive focus on test accuracy and interference from intrusive obsessive thoughts. This impairment may be associated with dysfunction in the ventral prefrontal cortex system [3]. Individuals with OCD may have concerns about incorrectly filling their exam numbers during exams, leading to repetitive checking behaviors that impact their cognitive reasoning and problem-solving skills, potentially resulting in an inability to finish the exams within the time constraints.

ERP is a form of Cognitive-Behavioral Therapy (CBT) designed to assist patients in confronting and exposing themselves to fears or discomfort related to their obsessive thoughts, with the goal of preventing the performance of compulsive behaviors to alleviate the distress. The ultimate goal of ERP (Exposure and Response Prevention) is to challenge patients' conditioned fear and response to stimuli, allowing them to experience the outcome and gain an understanding that the feared stimuli are, in fact, safe [4]. Amplifying the skin texture under certain circumstances exposes the patient to an immersive imperfect skin texture, which reduces the excessive focus on perfection and imperfection and acts as a palliative.

Trichotillomania (hair-pulling disorder) and skin-picking (excoriation) disorder are neuropsychiatric disorders that usually occur in conjunction with OCD, but are not recognized by professionals [5]. Skin-picking disorder is a psychiatric disorder characterized by repeated scratching or picking at the skin, resulting in skin injuries such as minor ulcers, hyperpigmentation, shallow scars, and even, less commonly, more severe skin disfigurement and skin infections [6–8]. Some of the characteristics of OCD can be explored by analyzing the textures of the area of skin injuries with the electromyographic information generated during the behaviors. In addition to medical consequences, psychological sequelae of skin scratching have been identified, including clinically significant distress and different areas of dysfunction [9]. Skin picking is categorized as a body-focused repetitive behavior (BFRB) characterized by recurrent and habitual actions directed at the body [10].

Trichotillomania is characterized by a repetitive act of pulling one's own hair, leading to hair loss and potential dysfunction [11]. This condition primarily involves pulling from the scalp, eyebrows, and eyelashes, although any body part with hair, such as the pubic area, can also be affected [12,13]. It is not uncommon for individuals with Trichotillomania to engage in hair pulling from multiple areas, and the episodes of pulling can vary in duration, ranging from a few minutes to several hours. Magnify and observe the multiple skin areas involved in the act of hair plucking, looking for specific skin features.

enabling the acquisition of a comprehensive dataset. Based on this data, the study draws conclusions regarding the unique manifestations of OCD symptoms related to human skin textures, providing novel insights into the intricate connections between these elements.

The innovation point of this study lies in its interdisciplinary approach, integrating cognitive and neurobiological aspects to analyze OCD through the lens of human skin textures. This holistic exploration of OCD-related behaviors and skin texture traces offers a novel perspective on the complexities of the disorder, potentially paving the way for advancements in research and treatment strategies. By unveiling potential links between OCD and skin textures, this study aims to contribute valuable insights to the broader understanding of the disorder, ultimately aiming to improve the lives of individuals affected by OCD.

2 Background

Obsessive-compulsive disorder (OCD) is a multidimensional heterogeneous disorder characterized by impairments in volitional processes, including attention, association, thinking, and behavioral autonomy. The core of this disorder lies in the disharmony within the self [1]. This phenomenon manifests as a conflict between rational thoughts and the intrusive, distressing obsessions characteristic of OCD. Patients may express frustration, guilt, and shame, recognizing the irrationality of their obsessions and compulsions but finding it exceedingly difficult to resist or control them. This internal turmoil can significantly impact their self-esteem and emotional well-being, contributing to a cycle of distress and compulsion. Addressing this disharmony becomes a central therapeutic goal, as it is vital for helping patients develop effective coping strategies and building resilience in managing OCD symptoms. Some individuals with OCD may present with comorbid dissociative identity disorder (DID), characterized by difficulties in memory judgment related to both actual and imagined execution. Patients exhibit a lack of confidence in their memories, resulting in compulsive symptoms such as repetitive checking [1]. Repeated checking behaviors include skin-to-object contact, such as touching the door handle when repeatedly checking whether the door is closed, rummaging through a bag when checking whether the contents are adequately stored, repeatedly touching the gas, water, and electricity switches with the fingers when checking whether the switches are turned off, and so on.

Due to the presence of memory impairments in OCD, patients tend to focus more on event details, affecting their memory function. Some scholars propose that episodic memory deficits are secondary to executive function impairment, attributed to deficits in memory encoding [2]. As a result of paying attention to detailed content, it is easy for patients to see only specific parts of things or events and become too obsessed with the content of those parts, thus becoming unable to control their thinking as well as their behavior, making it challenging to take a macroscopic view of things. For example, a perfectionist will not be able to tolerate the marks or imperfections on an object, and thus will not be

Keywords: Obsessive-compulsive disorder (OCD) · exposure and response prevention (ERP) therapy · excoriation disorder · chronic skin-picking · human skin textures · cognitive

1 Introduction

Obsessive-Compulsive Disorder (OCD) presents a complex and heterogeneous mental health challenge, requiring in-depth investigation to comprehend its underlying mechanisms fully. This paper ventures into the unexplored territory of OCD and its potential link to human skin textures, with a specific focus on Excoriation Disorder, commonly known as chronic skin-picking. By conducting a comprehensive analysis of existing literature, this study seeks to elucidate the cognitive aspects, memory impairments, and potential neurobiological factors contributing to this unique association. The theoretical framework draws upon Exposure and Response Prevention (ERP) therapy, which encourages patients to confront their fears and obsessions. An innovative approach utilizes skin texture immersion as a simulated ERP, exposing individuals to imperfections and targeting perfectionistic tendencies associated with OCD.

The exploration of OCD demands an understanding of its clinical manifestations, particularly compulsive repetitive movements that often involve the skin. Within the realm of OCD, memory disorders lead individuals to hyperfocus on event details, resulting in behaviors such as constantly enlarging objects, which, in turn, can influence memory function. A noteworthy aspect is the lasting traces of skin texture left on objects during repeated exposures, potentially offering insights into the connection between OCD behaviors and the skin itself. This background serves as a crucial foundation for comprehending the complexities of OCD and its manifestations related to human skin textures.

The motivation driving this study arises from the limitations encountered in existing approaches to understanding OCD fully. Traditional methodologies have faced challenges in exploring the intricate nuances of the disorder, necessitating innovative avenues for investigation. Focusing on human skin textures, particularly in the context of Excoriation Disorder, this paper aims to offer a fresh perspective on OCD analysis. The juxtaposition of this novel approach with conventional methods illuminates the potential benefits of using human skin textures to enhance our comprehension of OCD and its related behaviors. Additionally, this study addresses the shortcomings of previous approaches and underscores the necessity for innovative methodologies to unlock the intricate relationship between OCD and human skin textures.

Central to this thesis are the main ideas and methods, prominently featuring skin-based motivation. The study proposes the use of magnified skin texture details to simulate ERP therapy for OCD. This simulation endeavors to expose patients to imperfections, fostering a gradual reduction in perfectionistic tendencies. Amplifying skin texture to create an immersive environment for ERP therapy serves as a medium for personal interaction and bridges the gap between the individual and their surroundings. Data collection for this research encompasses diverse approaches, such as questionnaires and real-world observations,

Understanding Obsessive-Compulsive Disorder Through Human Skin Textures

Yazhen Zhu[1,2,3], Jian Chen[2,3,4], Yuwei Sun[5], and Wei Wang[2,3,6(✉)]

[1] Royal College of Art, London SW7 2EU, UK
[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
chenj589@mail2.sysu.edu.cn, ehomewang@ieee.org
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen, MSU-BITUniversity, Shenzhen 518172, Guangdong, China
[4] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518000, China
[5] Columbia University, New York, NY 10027, USA
ys3371@tc.columbia.edu
[6] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. Obsessive-Compulsive Disorder (OCD) is a complex and heterogeneous mental health condition that challenges our understanding of its underlying mechanisms. This paper explores the potential connection between OCD and human skin textures, particularly Excoriation Disorder (chronic skin-picking), through a comprehensive analysis of existing literature. Investigating cognitive aspects, memory impairments, and potential neurobiological factors contributing to this association, the study also examines the role of human-computer interaction (HCI) in data analysis and treatment approaches, with a focus on skin texture-related aspects. Additionally, the thesis delves into two entry points for understanding OCD through human skin texture. OCD's clinical manifestations involve compulsive repetitive movements, where memory disorders lead individuals to hyperfocus on event details, causing behaviors of constantly enlarging objects, leaving traces of skin texture on them. Drawing inspiration from Exposure and Response Prevention (ERP) therapy, the paper proposes magnifying skin texture details to simulate ERP, exposing patients to imperfections and reducing perfectionistic tendencies. Secondly, related OCD symptoms, like compulsive skin peeling, leave specific skin marks, providing potential clues for identifying OCD characteristics and patterns. This innovative approach offers valuable insights into the complexities of OCD, highlighting the significance of human skin texture in understanding and treating the disorder. By integrating cognitive and neurobiological aspects, this study provides a comprehensive perspective on the intriguing relationship between OCD and human skin textures, contributing to advancements in OCD research and intervention.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 85–95, 2024.
https://doi.org/10.1007/978-3-031-65126-7_9

11. Zakaria, C., Yilmaz, G., Mammen, P.M., Chee, M., Shenoy, P., Balan, R.: Sleepmore: inferring sleep duration at scale via multi-device wifi sensing. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **6**(4), 1–32 (2023)
12. Yu, B., et al.: Wifi-sleep: sleep stage monitoring using commodity wi-fi devices. IEEE Internet Things J. **8**(18), 13900–13913 (2021)
13. Yang, X., Yu, X., Xie, L., Xue, H., Zhou, M., Jiang, Q.: Sleep apnea monitoring system based on commodity wifi devices. Comput. Mater. Cont **2**(69), 2793–2806 (2021)
14. Liu, W., Chang, S., Liu, Y., Zhang, H.: Wi-PSG: detecting rhythmic movement disorder using cots wifi. IEEE Internet Things J. **8**(6), 4681–4696 (2020)
15. Ridolfi, M., Kaya, A., Berkvens, R., Weyn, M., Joseph, W., Poorter, E.D.: Self-calibration and collaborative localization for UWB positioning systems: a survey and future research directions. ACM Comput. Surv. (CSUR) **54**(4), 1–27 (2021)
16. Li, S., Wang, Z., Zhang, F., Jin, B.: Fine-grained respiration monitoring during overnight sleep using IR-UWB radar. In: Hara, T., Yamaguchi, H. (eds.) International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, pp. 84–101. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-94822-1_5
17. Han, W., Dai, S., Yuce, M.R.: Real-time contactless respiration monitoring from a radar sensor using image processing method. IEEE Sens. J. **22**(19), 19020–19029 (2022)
18. Kwon, H.B., et al.: Attention-based LSTM for non-contact sleep stage classification using IR-UWB radar. IEEE J. Biomed. Health Inform. **25**(10), 3844–3853 (2021)
19. Atlas, D., Srivastava, R., Sekhon, R.S.: Doppler radar characteristics of precipitation at vertical incidence. Rev. Geophys. **11**(1), 1–35 (1973)
20. Baboli, M., Singh, A., Soll, B., Boric-Lubecke, O., Lubecke, V.M.: Wireless sleep apnea detection using continuous wave quadrature doppler radar. IEEE Sens. J. **20**(1), 538–545 (2019)
21. Islam, S.M.M., Lubecke, V.M.: Sleep posture recognition with a dual-frequency microwave doppler radar and machine learning classifiers. IEEE Sensors Lett. **6**(3), 1–4 (2022)
22. Rahman, T., et al.: DoppleSleep: a contactless unobtrusive sleep sensing system using short-range doppler radar. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 39–50 (2015)
23. Chen, L., Xiong, J., Chen, X., Lee, S.I., Zhang, D., Yan, T., Fang, D.: Lungtrack: towards contactless and zero dead-zone respiration monitoring with commodity RFIDS. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(3), 1–22 (2019)
24. Liu, C., Xiong, J., Cai, L., Feng, L., Chen, X., Fang, D.: Beyond respiration: contactless sleep sound-activity recognition using RF signals. Proc. ACM Interact. Mob. Wearable Ubiquit. Technol. **3**(3), 1–22 (2019)
25. Zhao, M., Yue, S., Katabi, D., Jaakkola, T.S., Bianchi, M.T.: Learning sleep stages from radio signals: a conditional adversarial architecture. In: International Conference on Machine Learning, pp. 4100–4109. PMLR (2017)

In summary, each technology has its own merits and considerations. Contactless sensing also leaves much to be desired, such as greater noise immunity to the varying light conditions of different indoor environments. At the same time, because contactless sensing can capture more information, it faces more serious privacy issues. The choice depends on specific requirements, budget constraints, and the desired level of monitoring accuracy. Besides, more research can focus on how to combine these two methods for better performance and less cost.

5 Conclusion

In this work, we review the existing sleep monitoring methods based on Wifi sensors and wireless sensors. Then we make a comparative analysis between these two methods for a better illustration of wireless sensors used in the field of sleep monitoring. Through the summary of the existing methods, we can better find the direction for the follow-up research. However, in addition to wifi sensors and radar, acoustic and optical sensors are also beginning to be used in this field. Therefore, it is our future work to further summarize and analyze the advantages and disadvantages of these methods.

References

1. Perez-Pozuelo, I., et al.: The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ Digital Med. **3**(1), 42 (2020)
2. Bertisch, S.M., et al.: Insomnia with objective short sleep duration and risk of incident cardiovascular disease and all-cause mortality: sleep heart health study. Sleep 41(6), zsy047 (2018)
3. Zhou, Q., Zhang, M., Hu, D.: Dose-response association between sleep duration and obesity risk: a systematic review and meta-analysis of prospective cohort studies. Sleep Breathing **23**, 1035–1045 (2019)
4. Palagini, L., Hertenstein, E., Riemann, D., Nissen, C.: Sleep, insomnia and mental health. J. Sleep Res. **31**(4), e13628 (2022)
5. Rundo, J.V., Downey, R., III.: Polysomnography. Handb. Clin. Neurol. **160**, 381–392 (2019)
6. Engstrøm, M., Rugland, E., Heier, M.S.: Polysomnography (PSG) for studying sleep disorders. Tidsskrift for den Norske laegeforening: tidsskrift for praktisk medicin, ny raekke **133**(1), 58–62 (2013)
7. Rottenberg, F., Nguyen, T.-H., Dricot, J.-M., Horlin, F., Louveaux, J.: CSI-based versus RSS-based secret-key generation under correlated eavesdropping. IEEE Trans. Commun. **69**(3), 1868–1881 (2020)
8. Chen, Z., Zhang, L., Jiang, C., Cao, Z., Cui, W.: Wifi CSI based passive human activity recognition using attention based BLSTM. IEEE Trans. Mob. Comput. **18**(11), 2714–2724 (2018)
9. Gui, L., Ma, C., Sheng, B., Guo, Z., Cai, J., Xiao, F.: In-home monitoring sleep turnover activities and breath rate via wifi signals. IEEE Syst. J. **17**, 2355–2365 (2022)
10. Liu, J., Chen, Y., Wang, Y., Chen, X., Cheng, J., Yang, J.: Monitoring vital signs and postures during sleep using wifi signals. IEEE Internet Things J. **5**(3), 2071–2084 (2018)

Doppler radar is widely used in the field of sleep detection due to its excellent ability to measure target displacement remotely. Doppler radar can capture the information of chest displacement due to respiration or heartbeat through the transmitted microwave signals and analyze it through the Doppler effect [19]. A contactless system named PRMS using quadrature microwave doppler radar to monitor sleep apnea events in real time. The system contains a real-time actigraphy and sleep apnea detection algorithm [20]. A novel sleep posture recognition technique is proposed, which employs classifiers that are amenable to optimization through Bayesian hyperparameter tuning. These classifiers operate on data from a dual-frequency monostatic continuous-wave radar system [21]. DopplerSleep, a contact sleep sensing system, uses a single Doppler sensor to track sleep quality. DopplerSleep can monitor both body movements and tiny chest and heart movements, and the system has been experimentally validated to perform well on sleep stage classification tasks [22].

RF signals are widely used for contactless motion and vital signs monitoring in the field of sleep monitoring. Radio Frequency Identification (RFID) is a contactless communication technology that enables two-way data exchange for identification and data transfer using RF signals with flexibility and low cost. A respiration monitoring system with RFID sensors called LungTrack is proposed to achieve dual objective monitoring with an accuracy of above 93% for two targets at a distance of 10 cm at least [23]. TagSleep is a sleep posture recognition system using the concept of two-layer sensing with RFID sensors [24]. A model combining a convolutional network and recurrent neural network is trained on the RF-measured sleep dataset with an adversarial training regime [25].

4 Comparative Analysis

WiFi sensors and other wireless sensors, as non-interference devices, offer both advantages and disadvantages in sleep monitoring. Figure 1 shows a comparison between these two methods. WiFi sensors typically utilize wireless signals and receivers to track variables such as breathing, body movement, and sleeping positions. These sensors analyze movement patterns and breathing rates by observing changes in WiFi signals. They are cost-effective and easy to deploy, but privacy concerns may arise.

On the other hand, radar technology emits high-frequency pulse signals and measures the time it takes for the signals to bounce back. This enables accurate positioning and tracking of objects, including monitoring human movements and breathing patterns during sleep. Radar provides precise distance and position measurements, boasting high accuracy and reliability. However, radar requires specialized hardware and incurs higher costs. Both UWB and doppler radars described previously are capable of real-time sleep monitoring with a high degree of accuracy, but there is the problem of higher equipment costs and more demanding deployment conditions during equipment placement.

While RFID technology offers advantages like low power consumption and affordability, it may have limitations when it comes to more detailed sleep analysis and breathing monitoring.

wifi sensors are used for obstructive sleep apnea (OSA) detection and rhythmic movement disorder (RMD) detection. An intelligent apnea monitoring system can utilizing linear fitting and wavelet transform eliminate the phase error of CSI. The system uses commodity wifi, which is better able to eliminate interference from changes in sleeping posture [13]. A sleep monitoring system named Wi-PSG is proposed to utilize CSI from Wifi infrastructures for RMD-related movement detection, which can achieve an accuracy of above 92% for different RMD movement classifications [14].

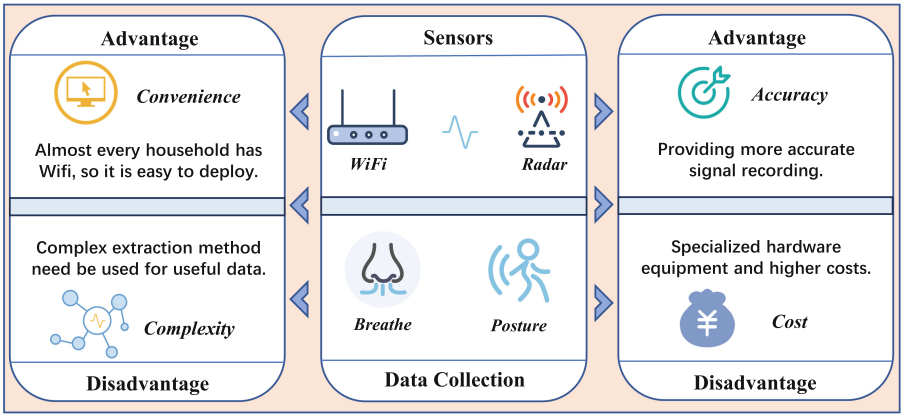


Fig. 1. Comparison between WiFi sensors and radars.

3 Sleep Detection Based on Wireless Sensor

Wireless radars are the most widely used sensors in sleep detection based on wireless sensors. Systems with wireless sensors are usually used for vital signs detection during sleep and sleep quality detection. The main sensors used in these systems are Ultrawideband (UWB) radar, Doppler radar, and Radio Frequency (RF) sensors.

UWB radar is commonly utilized for precise localization, employing low energy levels for short-range and high-bandwidth communications across the radio spectrum [15]. The required sleep information can be extracted by the UWB radar sensor penetrating the clothes and quilt. A fine-grained prototype for overnight respiration monitoring is proposed by exploiting the complementarity between the amplitude and phase of the radar signal [16]. Four respiration patterns are recognized during overnight sleep in this method. Another image processing method converts the raw signals collected by the UWB radar into a 2-D heatmap image and then an image-processing algorithm is used to capture respiratory information for respiratory motion measure [17]. An attention-based LSTM model is proposed to use the vital signs detected remotely by an impulse-radio UWB radar for sleep stage classification [18].

[6], However, the recording of PSG always needs expensive equipment and keep lots of contact with the subjects' body which bring discomfort. These drawbacks make it unsuitable for daily life sleep monitoring.

With the development of information techniques, more and more wireless sensors are used for sleep monitoring. There are already a lot of wearable devices used for sleep monitoring, but they also face resistance because of the discomfort brought to subjects and instability during sleep. Contactless sensors can effectively address the problem that invasive sensors bring natural sleep difficulties. There are various contactless sensors used in sleep monitoring now. The main of them are wireless sensors. Wifi sensor is also a kind of wireless sensor but since it has received more attention than other wireless sensors, it is put in a separate category.

Since wireless sensors are now widely used in sleep monitoring and have shown great potential, it is meaningful to review sleep monitoring research based on wifi sensors and wireless sensors. This can help develop contactless devices to achieve stable, safe, and non-contact sleep detection. In this work, we will first review the main sleep detection methods based on wifi sensors and wireless sensors respectively, and then a comparative analysis is made to summarize the difference between wifi sensors and wireless used in sleep monitoring. Finally, we provide a conclusion of our work.

2 Sleep Detection Based on Wifi Sensor

Wifi-based sleep monitoring activities are generally carried out through high precision indoor positioning, and the commonly used methods include Received Signal Strength (RSS) and Received Signal Strength (CSI) [7]. With the development of the technology, the CSI technique has demonstrated greater stability and accuracy and has become the more mainstream method nowadays. While using wifi sensors for sleep monitoring, CSI can be used to capture the effect of sleep activity contained by the Wifi signals [8].

Existing methods that use Wifi sensors to monitor sleep quality include heart rate monitoring and respiration monitoring [9]. A method is proposed to track the breathing rate and heart rate during sleep with Wifi [10]. They exploit to utilize the fine-grained channel information of existing Wifi networks to extract the minute movements that come with breathing and heartbeats. Wifi network activity is also used in a sleep-tracking approach called SleepMore which utilizes machine learning methods [11]. SleepMore constructs a semi-personalized random forest model to make a classification of the network activity behavior and the results are divided into sleep and awake states in minute dimensions. The experimental results show that SleepMore achieves an indistinguishable result with the Oura ring baseline within a 5% uncertainty rate.

Wifi sensors are also used for sleep stage classification and sleep-related disorders detection. An advanced signal processing and fusion method is proposed to extract accurate respiration and body movement for four-stage sleep classification, which achieves an accuracy of 81.1% [12]. In disorders monitoring,

Review of Sleep Monitoring Research Based on Wireless Sensor

Yuzhu Hu[1,2,3], Jian Chen[1,2,3], Shen Zhao[1(✉)], Kexin Tan[2], Kuai Yu[2], and Wei Wang[2,3,4]

[1] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518000, China
{huyzh27,chenj589}@mail2.sysu.edu.cn, z-s-06@163.com
[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{1120200259,1120200296}@smbu.edu.cn, ehomewang@ieee.org
[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen MSU-BITUniversity, Shenzhen 518172, Guangdong, China
[4] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. Since sleep quality is crucial to human health, sleep monitoring has become a hot spot in the field of smart healthcare. Previous methods depend on polysomnography and wearable devices need immediate contact with the subject, which brings discomfort. Contactless sensors can address this issue. The most common contactless sensors used in sleep monitoring are wireless sensors (including radar and WiFi). To clarify the research in this area, we summarized the existing sleep monitoring methods based on WiFi sensors and wireless radar and made a comparison. The conclusion shows that the two kinds of methods have advantages and disadvantages, so the development of complementary methods is very promising for sleep monitoring.

Keywords: Sleep monitoring · contactless sensors · wireless sensing

1 Introduction

Sleep is one of the most important basic life activities of human beings, and it is also an important basis for maintaining physical and mental health [1]. Chronic poor sleep has also been linked to cardiovascular disease, obesity, and even some mental health problems [2–4]. Therefore, sleep monitoring is important for health status monitoring and is now become a hot topic for research.

Polysomnography (PSG) is the most widely used tool to monitor sleep, and it is regarded as the gold standard to detect sleep-related breathing disorders [5]. PSG can provide comprehensive information on sleep stages on the basis of Electroencephalography (EEG) activity, eye movements, and muscular tension

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 79–84, 2024.
https://doi.org/10.1007/978-3-031-65126-7_8

15. Vos, T., et al.: Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet **396**(10258), 1204–1222 (2020)
16. Walker, E.R., McGee, R.E., Druss, B.G.: Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry **72**(4), 334–341 (2015)

developing countries (Fig. 2); anxiety contributes to the development of depression and must be taken into account as well.

Considering all above we can suggest the authorities to take more measures to ease the burden and stress of the depraved people. As other studies showed [4,6,8], low-income group are at the higher risk of getting depression and having worse health condition in general [3,5], so, some government financial help is better be provided (subsidiaries, money allowance, etc.).

Acknowledgment. This work is supported by the Shenzhen Science and Technology Innovation Commission (Stabilisation Support Programme).

References

1. Chesney, E., Goodwin, G.M., Fazel, S.: Risks of all-cause and suicide mortality in mental disorders: a meta-review. World Psychiatry **13**(2), 153–160 (2014)
2. Cuijpers, P., Vogelzangs, N., Twisk, J., Kleiboer, A., Li, J., Penninx, B.W.: Comprehensive meta-analysis of excess mortality in depression in the general community versus patients with specific illnesses. Am. J. Psychiatry **171**(4), 453–462 (2014)
3. Diener, E., Biswas-Diener, R.: Will money increase subjective well-being? Soc. Indic. Res. **57**, 119–169 (2002)
4. Dwyer, R.J., Dunn, E.W.: Wealth redistribution promotes happiness. Proc. Natl. Acad. Sci. **119**(46), e2211123119 (2022)
5. Headey, B., Muffels, R., Wooden, M.: Money does not buy happiness: or does it? a reassessment based on the combined effects of wealth, income and consumption. Soc. Indic. Res. **87**, 65–82 (2008)
6. Kahneman, D., Deaton, A.: High income improves evaluation of life but not emotional well-being. Proc. Natl. Acad. Sci. **107**(38), 16489–16493 (2010)
7. Kartaev, P.S.: How to teach econometrics to economists: bachelor level..... 72 macroeconomic policy. Sci. Res. Fac. Econ. Electron. J. **11**(2), 72–90 (2019)
8. Killingsworth, M.A.: Experienced well-being rises with income, even above $75,000 per year. Proc. Natl. Acad. Sci. **118**(4), e2016976118 (2021)
9. Patel, V., et al.: Addressing the burden of mental, neurological, and substance use disorders: key messages from disease control priorities. The Lancet **387**(10028), 1672–1685 (2016)
10. Pearce, M., et al.: Association between physical activity and risk of depression: a systematic review and meta-analysis. JAMA Psychiatry **79**, 550–559 (2022)
11. Reger, M.A., Stanley, I.H., Joiner, T.E.: Suicide mortality and coronavirus disease 2019-a perfect storm? JAMA Psychiatry **77**(11), 1093–1094 (2020)
12. Son, J., Shin, J.: Bimodal effects of sunlight on major depressive disorder. Compr. Psychiatry **108**, 152232 (2021)
13. Stock, J.H., Watson, M.W.: Introduction to Econometrics, vol. 104. Addison Wesley Boston (2003)
14. Viswanathan, M., et al.: Screening for depression and suicide risk in children and adolescents: updated evidence report and systematic review for the us preventive services task force. JAMA **328**(15), 1543–1556 (2022)

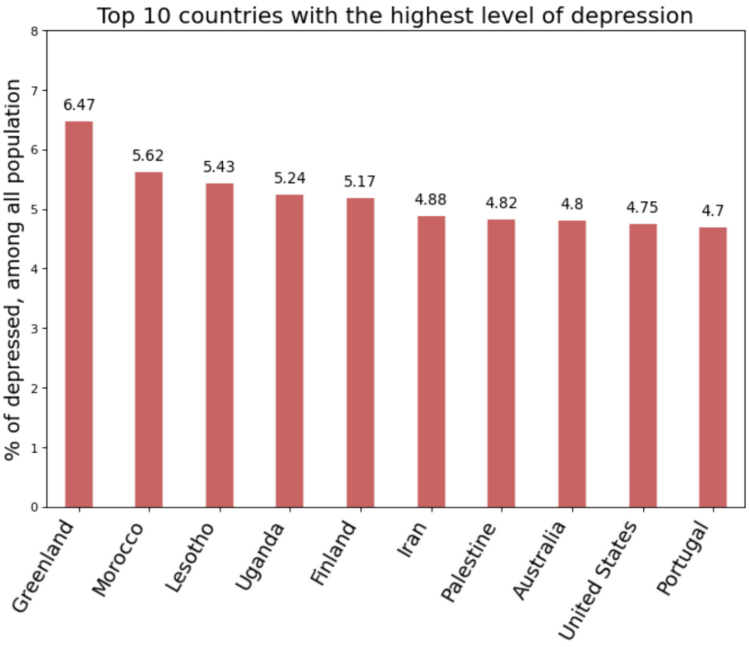


Fig. 2. Top 10 countries with the highest level of depression

As can be noticed, top ten includes mainly developing countries where GDP per capita quite small. The exceptions are Finland, Australia, United States, Portugal and Greenland as a special region of Denmark, where the GDP per capita is medium or higher. In case with Finland and Greenland, such higher level of depression could be explained by two factors: 1) isolated and low populated communities, as can be seen from the depression model, the population size is significant factor; 2) the lack of sunny days, what negatively effects on mood and emotional conditions [12]. As for other countries, the further deep analysis is required.

4 Discussion

This large-scale study based on worldwide panel data about depression showed that people who live in countries with low GDP per capita are more vulnerable to depression. We find that the relationship between depression and GDP per capita is strongly negative, and because of analyses of huge massive of date, the results are universal. At the same time, the connection between depression and anxiety disorders is strongly positive, thus, the following conclusions could be made: in countries with lower GDP per capita, more people tend to suffer from depression. Actually, this fact can be proved even statistically: the majority of the countries in the list of top 10 countries with high level of depression are

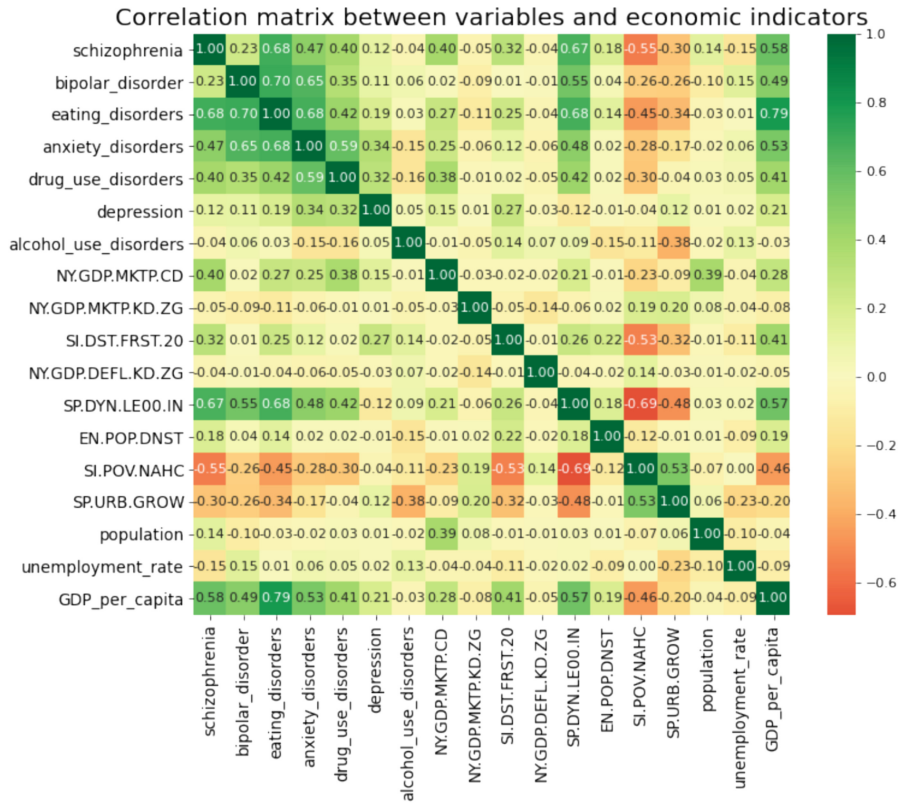


Fig. 1. Correlation matrix between variables and economic indicators

- GDP per capita - all things being equal, with an increase in GDP by one dollar, the number of people suffering from depression decreases by 2,74*e-6%
- Income share held by lowest 20% - all things being equal, with an increase in Income share held by lowest 20% by one dollar, the number of people suffering from depression decreases by 0,024%
- Life expectancy at birth, total - all things being equal, with an increase in life expectancy at birth, by one year, the number of people suffering from depression decreases by 0,013%
- Population - all things being equal, with an increase in population by one people, the number of people suffering from depression decreases by 1,05*e-9%
- Anxiety disorders - all things being equal, with an increase in anxiety disorders by one percent, the number of people suffering from depression increases by 0,35%

Now countries that have the highest rates of depression are shown in Fig. 2.

– SI.POV.NAHC - Poverty headcount ratio at national poverty lines (% of population)
– SP.URB.GROW - Urban population growth (annual %)
– Unemployment - Unemployment rate, (% of work force)
– GDP_PER_CAPITA - GDP per capita, (current US$)

2.3 Panel Study

In order to avoid omitted variable bias, we took regressors from different spheres [7,13]: pure economic, social and urban. We also have added the variable of control - Anxiety, as, by all means, anxiety disorders influence on the development of depression and other mental disorders. We conducted all measures using special econometric program Gretl.

3 Results

We calculated the correlation between all mental disorders and economic indicators as Fig. 1.

At the same time, we got the following depression model as Tabel 1.

Table 1. Depression Model

	Coefficient	St. error	t-statistics	p-value	
const	325,388	0,707587	4,599	¡0,0001	***
anxiety disorders	0,354180	0,145535	2,434	0,0168	**
NYGDPMKTPCD	0,000000	0,000000	4,352	¡0,0001	***
NYGDPMKTPKDZG	0,00112238	0,00100377	1,118	0,2663	
SIDSTFRST20	−0,0239324	0,00750013	−3,191	0,0019	***
NYGDPDEFLKDZG	−0,000230614	0,000412541	−0,5590	0,5775	
SPDYNLE00IN	−0,0134092	0,00500963	−2,677	0,0088	***
ENPOPDNST	0,000163380	0,000295864	0,5522	0,5821	
SIPOVNAHC	−0,00107081	0,00145502	−0,7359	0,4636	
SPURBGROW	−0,00724004	0,00900659	−0,8039	0,4235	
population	−1,04580e−09	1.73E−05	−6,051	< 0,0001	***
unemployment_rate	−0,00358409	0,00199508	−1,796	0,0756	*
GDP_PER_CAPITA	−2,74364e−06	9.21E−02	−2,978	0,0037	***

The LSDV R-square for this model is 0,9956, '*' means that variable is significant on 10%, '**' - 5%, and '***' - 1%. Therefore, we could interpret four variables of interest (on 5%):

World Health Association, around 280 million of people worldwide are suffering from depression, moreover, the World Health Organization assumes that 5% of men and 9% of female experience depressive disorders in their lifetime [10, 15]. Depression can lead to the development of other illnesses what effect on premature mortality [1,2,16] and even increase the suicide rates [9,11,14], that is why it is crucial for authorities to be aware of development of such illnesses. The innovation of this work is that it includes factors and figures from different spheres and examine their impact on the development of depression and other mental disorders. This allows us to broaden our thinking and to make more clear judgments [7,13]. Particularly, in addition to social-economic indicators, we also added urban population growth in our list of economic indicators, what allows to see the big picture. This article is aimed to determine how the main economic indicators are connected with mental disorders. After establishing the relationships, it will be possible to judge whether the country at the risk of mass depression. We believe that with the help of our research local authorities will be able to identify the upcoming health threats more effectively, and, what is the key point, much earlier, thus, many human lives would be improved or even saved.

2 Methods

This is a panel study which includes data from 196 countries throughout 27 years. In our research we mainly used econometrics and ordinary least square (OLS) analysis to make proper models. All implemented models have passed the Ramsey Test, the check for heteroscedasticity and multicollinearity, thus, all described models are trustful. Besides, in case with the depression analysis, the Fixed Effects model was used due to take into account each country peculiarity [7,13].

2.1 Dependent Variables

In addition to Depression, we also considered the following types of mental diseases: Schizophrenia, Bipolar disorder, Eating disorders, Anxiety disorders, Drug use disorders, Alcohol use disorders. All variables are examined as % of all population.

2.2 Economic Indicators

For each variable we make an econometric model with the following regressors:

- NY.GDP.MKTP.CD - GDP (current US$)
- NY.GDP.MKTP.KD.ZG - GDP growth (annual %)
- SI.DST.FRST.20 - Income share held by lowest 20%
- NY.GDP.DEFL.KD.ZG - Inflation, GDP deflator (annual %)
- SP.DYN.LE00.IN - Life expectancy at birth, total (years)
- EN.POP.DNST - Population density (people per sq. km of land area)

Identification of Economic Factors for Mass Depression Based on Panel Study and Machine Learning

Iaroslava Pravolamskaya[1,2,3,4], Jian Chen[3,4,5], and Wei Wang[3,4,6(✉)]

[1] Faculty of Economics, Shenzhen MSU-BIT University, Shenzhen 518172, China
[2] Faculty of Economics, Lomonosov Moscow State University, Moscow 119991, Russia
[3] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
[4] Guangdong-Hong Kong-Macao Joint Laboratory for Emotion Intelligence and Pervasive Computing, Shenzhen, MSU-BITUniversity, Shenzhen 518172, Guangdong, China
[5] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518000, China
chenj589@mail2.sysu.edu.cn
[6] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
ehomewang@ieee.org

Abstract. Panel study and machine learning are important tools for analyzing various aspects of the economy. They allow researchers to study the dynamics of changes in different economic indicators, such as GDP, inflation, unemployment, etc. In addition, these tools can be used to determine causal relationships between social, economic and psychological factors what can allow us to predict the development of the economy and changes in people's life in the future. However, previous works in this sphere studied the connections between income and happiness, not taking into account the relationships between economic indicators and mental disorders. This article is aimed to analyze the relationship between economic factors and the level of mass depression based on a panel study and machine learning methods. Experimental results based on panel study and machine learning demonstrate effectiveness of our proposed econometric model.

Keywords: Panel Study Machine Learning Depression Identification Economic Factors Econometrics Models

1 Introduction

The increasing number of people suffering from depression and other mental diseases is one of the most challenging issues in the 21 century. According to

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 72–78, 2024.
https://doi.org/10.1007/978-3-031-65126-7_7

14. Koyama, K., Hoshikawa, H., Kojima, G.: Eddy current nondestructive testing for carbon fiber-reinforced composites. J. Press. Vessel. Technol. **135**(4), 041501 (2013)
15. Kostopoulos, V., Vavouliotis, A., Karapappas, P., Tsotra, P., Paipetis, A.: Damage monitoring of carbon fiber reinforced laminates using resistance measurements. Improving sensitivity using carbon nanotube doped epoxy matrix system. J. Intell. Mater. Syst. Struct. **20**(9), 1025–1034 (2009). https://doi.org/10.1177/1045389X08099993

conducted to get the correct data and then compared to draw conclusions. Damage to the carbon/glass blend and fracture of the carbon fibers was observed by using Three Point Bending method. The pictures show that where pressure is applied the upper layers are damaged by shear stresses leading to kinking and the lower layers are damaged mainly in the form of delamination leading to failure.

In conclusion, this study has been designed, experimented and concluded that it is feasible to monitor the electrical conductivity of this hybrid carbon/glass fiber blend and that this composite fiber can also be seen as a self-sensor.

Acknowledgments. This project is supported by the funding of Guangdong Province Key Laboratory of Intelligent Detection in Complex Environment of Aerospace, Land and Sea. (2022KSYS016).

References

1. Maleque, M.A., Salit, M.S.: Materials Selection and Design. SM, Springer, Singapore (2013). https://doi.org/10.1007/978-981-4560-38-2
2. Jalalvand, M., Czél, G., Wisnom, M.R.: Damage analysis of pseudo-ductile thin-ply UD hybrid composites – A new analytical method. Compos. A Appl. Sci. Manuf. **69**, 83–93 (2015). https://doi.org/10.1016/j.compositesa.2014.11.006
3. Sauer, M.: Composites Market Report 2019—The Global CF-und CC-Market 2019: Market Developments, Trends, Outlook and Challenges. Composites United eV, Berlin, Deutschland (2019)
4. Czél, G., Wisnom, M.R.: Demonstration of pseudo-ductility in high performance glass/epoxy composites by hybridization with thin-ply carbon prepreg. Compos. A Appl. Sci. Manuf. **52**, 23–30 (2013)
5. David-West, O., et al.: A review of structural health monitoring techniques as applied to composite structures. Struct. Durability Health Monit. **11**(2), 91–147 (2017)
6. Rev, T., et al.: A simple and robust approach for visual overload indication-UD thin-ply hybrid composite sensors. Compos. A Appl. Sci. Manuf. **121**, 376–385 (2019)
7. Chapuis, B.: Introduction to structural health monitoring. In: Chapuis, B., Sjerve, E. (eds.) Sensors, Algorithms and Applications for Structural Health Monitoring. IC, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-69233-3_1
8. Vavouliotis, A., Paipetis, A., Kostopoulos, V.: On the fatigue life prediction of CFRP laminates using the electrical resistance change method. Compos. Sci. Technol. **1**(5), 630–642 (2011)
9. Smith, R.: Composite defects and their detection. Mater. Sci. Eng. **3**(1), 103–143 (2009)
10. Gregor Trtnik, M.G.: Recent advances of ultrasonic testing of cement based materials at early ages. Ultrasonics **54**, 66–75 (2013)
11. Song, S., Jing, J., Cheng, W.: Online monitoring system for macro-fatigue characteristics of glass fiber composite materials based on machine vision. IEEE Trans. Instrum. Meas. **71**, 1–12 (2022)
12. Manjunatha, P.A.: Vision-based and data-driven analytical and experimental studies into condition assessment and change detection of evolving civil, mechanical and aerospace infrastructures. Doctoral dissertation, University of Southern California (2022)
13. Bayraktar, E., Antolovich, S.D., Bathias, C.: New developments in non-destructive controls of the composite materials and applications in manufacturing engineering. J. Mater. Process. Technol. **206**(1–3), 30–44 (2008). https://doi.org/10.1016/j.jmatprotec.2007.12.001

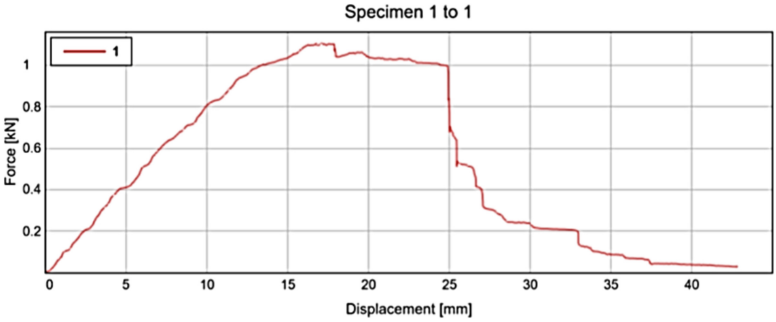


Fig. 11. 10 mm experimental group displacement and Force curve

to the phenomenon of pseudo-stretchability. Analysis of the sample damage showed that shear damage dominated at the upper end of the sample, while delamination dominated from the middle to the lower plies, shown as Fig. 12. However, the main change in resistance in this test was due to the fracture of the thin carbon fibers, which was mainly due to tensile stresses, while the delamination of the lower plies was mainly due to shear stresses, so the design of this test was reasonable.

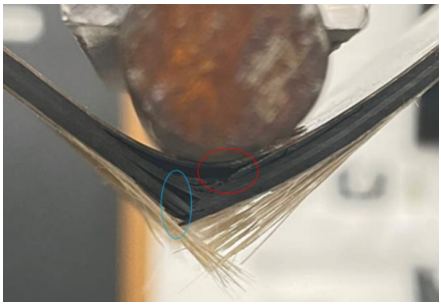


Fig. 12. Injury patterns under three -point bending

6 Conclusion

In this project, a hybrid thin-layer carbon/glass fiber self-sensing method is proposed. It is innovative in that it changes the traditional case of applying the carbon fibers directly to the object to be sensed. Also, by using an S -shape instead of the traditional direct strip, it allows for greater coverage and a larger area to be monitored than just partial detection, while its more holistic nature makes it more effective for monitoring a whole plane rather than monitoring a broken location, and also has a greater improvement in monitoring the effects of certain unseen damage. Regarding the experiment, this experiment uses the controlled variable method to create differences for different variables. By designing groups of different widths as well as different styles, several experiments were

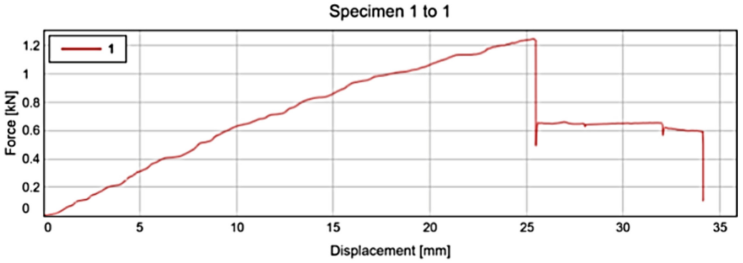


Fig. 9. 5 mm experimental group displacement and Force curve

in the 25–30 mm zone, the resistance begins to rise, indicating that the carbon fiber body is further destroyed in this zone, when in the 30–35 zone, the carbon fiber is completely destroyed and the resistance rises to 4000 Ω (Fig. 8). When the carbon fibers are completely destroyed, the loading force is removed and the fibers spring back, at this time some of the fibers reduce in resistance because the stress is reunited (Fig. 9).

10 mm Bending Test

The 10 mm three-point bending test is also primarily a comparison with 5 mm, observing the change in resistance of two different widths of carbon fiber to determine which is more appropriate.

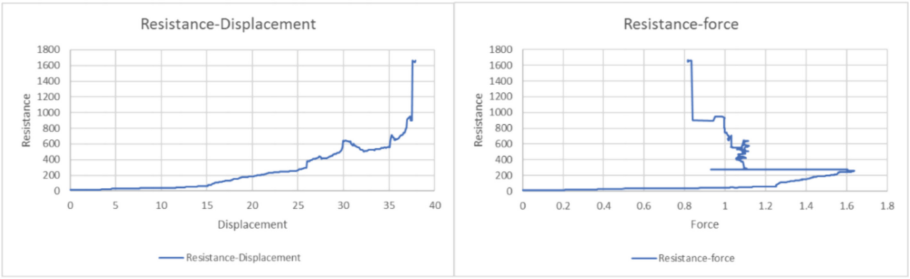


Fig. 10. 10 mm experimental group resistance displacement curve(left) and 10 mm experimental group resistance Force curve(right)

By comparing the two sets of plots, it can be found that the 5 mm images of resistance and Force are relatively similar to the 10 mm images on the three -point bending method, but on the resistance displacement curve, it is obvious that the rising trend of the 10 mm curve is smoother, so after the comparison, it is more recommended to use a 10 mm wide thin layerof carbon fiber as a self-sensor (Figs. 10 and 11).

Damage Mode Analysis

In this section, the damage pattern of the experimental product and the image in the above figure will be analyzed in detail, as the damage to the sample occurred gradually over the course of the test and this fiber hybridization slowed down the catastrophic rate and so led

Table 1. Mechanical properties of curing

Properties	Numerical value	Unit
Tg Onset(DMA)	**140**	°C
Tensile Strength	**645**	MPa
Compressive Strength	**515**	MPa
Flexural Strength	**882**	MPa
Flexural Modulus	**60.1**	GPa
Interlaminar Shear Strength	**69.8**	MPa
Tg Peak(DMA)	**148**	°C

of plain carbon fiber strips alone. As the fiber orientation was also considered in this carbon fiber experiment to affect the magnitude of the current, a unidirectional thin layer of carbon fiber was used in this case so that the consistency of the current could be maintained throughout.

In the three-point bending test, since the three -point bending test causes large shear stresses, data were collected from the start to sample failure and finally the changes in resistance and the reasons for these changes were analyzed in conjunction with the changes in the curves.

5 mm Bending Test

In this section the experimental data on the 5 mm three-point bending method is described. Unlike the above, as this design is a hybrid design, the standard T700 carbon fiber bending performance criteria above can only be used as a reference value, so according to the experimental process, the bending performance is significantly lower compared to T700, only around 800 Mpa, so it is speculated that it is possible that the mixture of glass fiber and thin-layered carbon fiber has affected the bending performance.

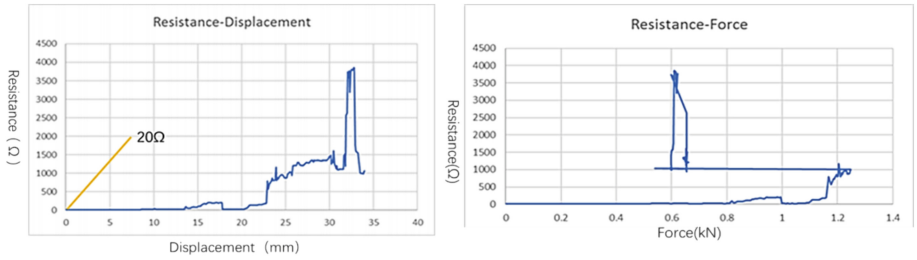


Fig. 8. 5 mm experimental group resistance displacement curve(top) and 5 mm experimental group resistance Force curve(bottom)

According to the data we can see that there is a relatively obvious increase in resistance after the indentation test, as can be seen from the graph, at 25 mm of the experiment is the maximum stress, when the carbon fiber begins to destroy, it can be concluded that

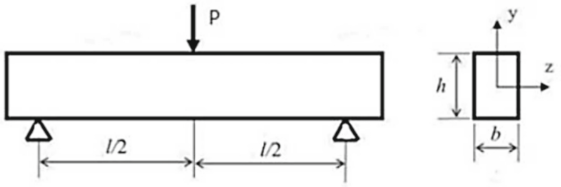


Fig. 7. Three-point bending method model

In the three-point bending test, when viewed from the front, it can simply be seen as a simply supported beam subjected to a concentrated pressure. Three-point bending should theoretically result in a linear distribution of positive stresses along the beam in the cross-sectional area when loaded.

$$\sigma = \frac{M}{I_z} y \tag{1}$$

where σ is the stress, M is the moment, Iz is the moment of inertia of the cross-section to the z-axis and y is the distance in the cross-section to the y-axis. The maximum positive stress at the danger point of the beam is:

$$\sigma_{Max} = \frac{M_{Max}}{I_z} y_{Max} \tag{2}$$

For rectangular section specimens:

$$M = \frac{P \times l}{4} \qquad I_z = \frac{bh^3}{12} \tag{3}$$

Substituting Eqs. (3) into (2) yields the new equation

$$\sigma_{bb} = \frac{3P \times l}{2bh^2} \tag{4}$$

where P is the load and L is the span, b is for width, h is for thickness.

In the case of a sample based on this equation, the maximum shear stress is calculated a

$$s\sigma_{bb} = \frac{3P \times l}{2bh^2} = \frac{3 \times 1.5\,KN \times 150\,mm}{2 \times 50\,mm \times 9\,mm^2} = 800\,Mpa \tag{5}$$

According to the Table 1, its standard value is 880 Mpa, However, as this design contains other fibers of different thicknesses or patterns, this data can only be used as a reference value for the main body of the sample, so in principle the maximum acceptable shear stress for this design should be lower than this value.

5 Results

This experiment focused on the fabrication process of the self-sensor, which was designed using an innovative mixture of carbon fiber and thin layers of E glass fiber, and investigated the advantages and differences between this combination and the use

widely used in destructive testing due to its simple construction and the fact that it does not require much manipulation. For this test, the sample is placed on a jig and a multimeter is connected to the two sections of copper to read the resistance data. The movement speed of 7 mm/min is entered into the control of the hydraulic press and the test is started (Fig. 4).

Fig. 4. Three -point bending test.

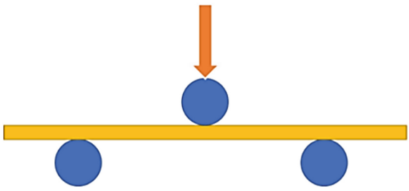


Fig. 5. Injury patterns under three-point bending

The three-point bending process is achieved mainly by applying pressure to the tip, during which the sample undergoes a process of gradual destruction. The following diagram shows the principle of three -point bending (Fig. 5).

During the three-point bending experiment, the sample started to break gradually when it was loaded to a high enough stress. As this sample was a mixed sample, the surface glass fiber started to break when it was loaded to 0.7 KN, the glass fiber broke completely when it was loaded to 1.2 KN, then the load was reduced to 0.6 KN and then the carbon fiber started to break gradually.

The images show that the entire damage process is produced gradually, and based on the experimental images it can be seen that the samples start with damage and end up with damage (Fig. 6).

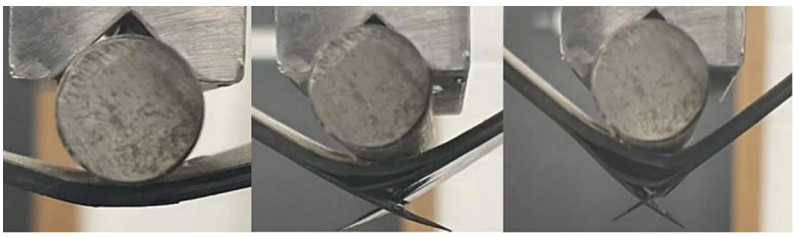


Fig. 6. The process of three-point bending test

Theory of Three-Point Bending Test

When bending deformation occurs in the three-point bending method, the fibers near the bottom elongate and those near the top shorten. According to the planar hypothesis, the fiber state changes gradually from stretching to compression along the height of the cross section from the bottom to the top, then there must be a layer in between where the length of the fiber remains constant, this layer is called the neutral layer (Fig. 7).

known as multi -directional (MD) fibers (Fig. 2). This multi -axial material has better tensile and compressive resistance than uniaxial material, but because it is manufactured at an angle, it is less malleable.

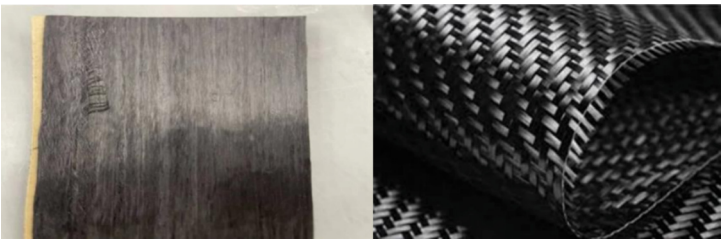


Fig. 2. UD carbon fiber(A), MD carbon fiber(B)

Experiment Test Group

The carbon fibers in the experimental group will be linked to each other, showing S-shaped connections, which then means that the data from the experimental group will affect each other. This control group can be used to see if the resistance will be affected by the occurrence of fiber breaks. Having established that the resistance will change due to fiber breakage, then this experimental group has the advantage that only two electrodes are needed to complete the experiment due to the large area it covers. The main reason for using two different widths of samples was to see the rate of change in resistance by comparing the two sizes of 5 mm and 10 mm. In the graph below, sample 1 is the control group of 10 mm, sample 2 is the control group of 5 mm, sample 3 is the experimental group of 5 mm and sample 4 is the experimental group of 10 mm, shown as Fig. 3.

Fig. 3. Kinds of simple.

Three-Point Bending Test

After the indentation test, the material is tested using the three-point bending method, which is one of the simplest and most effective methods of testing laminates and is

4 Research Methodology

It describes the methods and techniques used in the selection of materials, the preparation process, and the design of the testing process for this project. In particular, first, it will be described in detail what materials are used to prepare the samples, as well as the preparation process and methods. Then, it will be described the test process of the experiment and other equipment used in the experimental process, last, it is going to be described the detail of the whole experiment and the theoretical data will be given, including the theoretical currents and the theoretical stresses generated by the experiment.

Material Selection

In addition, for sample preparation we used T700/XC130 unidirectional prepreg carbon fiber as the sample body and S -Glass/913 and M46JB unidirectional prepreg thin carbon fiber as the sensor. The base material was made from T700 carbon fiber manufactured by Toray of Japan. When selecting the substrate material, it was considered that the main carbon fibers are mainly T300 and T700, both of which contain a large amount of carbon, but the overall performance of T700 is significantly better than that of T300. In the selection of the sensing layer, we chose to use a thin carbon fiber sandwiched between the two glass fibers. As the thin carbon fiber chosen, M46JB, has a similar tensile strength to T700, but obviously the compressive strength of M46JB is weaker than that of T700.T In this experiment, glass fibers were chosen to wrap the thin carbon fiber because, as seen in Meisam's model, there are three different damage modes for composites made from high and low strain materials, so in order for the sensing layer of carbon fibers to break before the glass fibers in the isolation layer, a thin layer of carbon fibers with a lower degree of strain than the glass fibers must be used as the induction material (Fig. 1). (Fotouhi, Jalalvand et al., 2017)

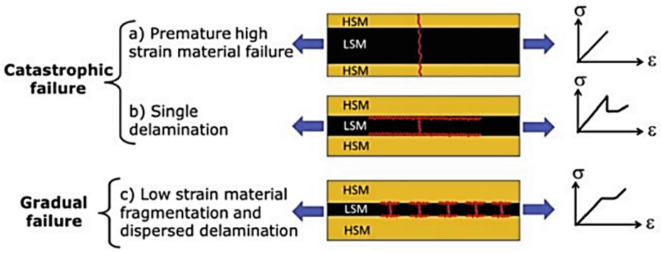


Fig. 1. Possible failure modes in a three layers UD hybrid made from HSM and LSM (red lines show fracture) (a) single crack through the whole specimen, (b) single crack in the LSM followed by instantaneous delamination, and (c) multiple fracture and localised stable pull-out of the LSM

Typically, the basic constituent material of a carbon fiber reinforced material is usually a combination of multiple or unidirectional fiber orientations, rearranged to provide different mechanical properties. A single unidirectional (UD) fiber arrangement, where all the fibers in the resin are aligned in one direction with no voids or breaks. Another type of arrangement is where the fibers are aligned at 0 and 90 degrees, this is

In fields such as aviation and construction, it is important to ensure safety margins, as sudden damage can potentially lead to injury or death as well as huge financial losses. In these important areas, sudden failures as well as small residual load capacities are not allowed. This is why higher safety margins and more conservative structural designs are the dominant design approach in current designs, while another problem with engineered materials is that they break down without prior detectable damage and warning [2].

2 Literature Review

The composite material is made by two or more components. Composite material can be made by using fiber that are cured within a resin. Using a combination of different materials, taking advantage of their strengths and reducing the impact of their weaknesses is an important idea in designing composite materials. The most common types are combining carbon fibers and glass fibers with a thermosetting resin to create either a CFRPs or GFRPs. The use of multiple fiber-reinforced polymer laminations to form a new composite material with enhanced mechanical properties, such as compressive and tensile resistance, and the consideration of how to efficiently monitor the new composite material, based on previous research by scientists, has become an important key, and finally some of the advantages and disadvantages of previous monitoring methods in relation to the composite material in this experiment are presented.

3 Aims and Objective

How to monitor the health of composite materials is an important current research issue. It can effectively contribute to the development of several areas where composites are used, such as aerospace engineering and the automotive industry. However, in the traditional composite fiber industry the damage itself is unpredictable and sudden as it can be caused by different kind of stresses and pseudo-plastic deformations. There are many different methods of detection, but most of them do not reflect the changes in the material in real time and are also too expensive. This project therefore proposes a method to monitor changes in material resistance based on the fact that the resistance of thin carbon fibers changes gradually during damage. Research done by Meisam has shown that damage to fibers varies linearly (Rev, Jalalvand et al. 2019) and therefore the health of the material can be monitored by detecting changes in resistance based on this theory. The aim of this project is to design another sensor based on resistance variation, to experiment in order to check if the sensor is usable, and to design and test the sensor in order to achieve the best possible results, observing through experimentation and results whether it is sensitive and efficient.

Research Objective 1. Research test samples and target sensors based on existing literature and conjecture.

Research Objective 2. Analysis of the change in resistance of the sample and its force and displacement profiles.

Research Objective 3. Using a hybrid S-shape to detect damage on a flat plane and resistance changes monitored using a carbon glass hybrid.

Structural Health Monitoring of Carbon Fiber Composite Lamination Using Electrical Resistance

Guiping Lu[1(✉)], Xiaofeng Zhang[1], Shan Lu[2], Binghua Su[1], Kejun Wang[1], and Jiaran Liang[1]

[1] Beijing Institute of Technology, Zhuhai, Zhuhai, China
344088386@qq.com
[2] BMW Brilliance Automotive Ltd, Shenyang, China

Abstract. It focuses on a composite material made of glass and carbon fibers in this paper. The composite can be actively monitored and controlled by the self-sensing of the carbon fibers. However, due to the high stiffness and brittleness of the composite material, damage often occurs instantaneously. It is difficult to monitor damage patterns and control damage through factors such as fiber type variables and displacement relationships. This is why monitoring the health of composite fibers is an important direction, which has major implications for the aerospace, industrial and automotive sectors. In this project, the main focus is to monitor the electrical conductivity of carbon fibers online by breaking thin layers and observing the changes in their conductivity, and to understand changes in condition through changes in current. In addition, the composite design of this project can be applied to the monitoring of large planar materials, as well as to applications in important areas such as aerospace. In making further comparisons, it can be seen that the 5 mm thin layer of carbon fiber is more sensitive in the process of self-sensing, while the change in resistance is more noticeable when damage is received in the period.

Keywords: Structural Health Monitoring · Carbon Fiber Composite Lamination · Electrical Resistance · Three-point bending test

1 Introduction

There are more and more high-tech products made of composite materials, such as aerospace or automotive, and even the latest batteries. They are becoming increasingly popular due to their outstanding properties, such as their high strength, low weight and fatigue resistance, Composites are combinations of two or more materials with different physical behavior and chemical states. In particular in this test, the materials used are fiber reinforced polymers. As the properties of composites are usually more variable, engineers consider their design structure and components to minimize failure during the design of composites [1].

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 61–71, 2024.
https://doi.org/10.1007/978-3-031-65126-7_6

5. Carrard, V., et al.: Medical student mental health. https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health
6. Davis, C., Martin, G., Kosky, R., O'Hanlon, A.: Early intervention in the mental health of young people: a literature review. In: ERIC (2000)
7. Ediz, B., Ozcakir, A., Bilgel, N.: Depression and anxiety among medical students: examining scores of the beck depression and anxiety inventory and the depression anxiety and stress scale with student characteristics. Cogent Psychol. **4**(1), 1283829 (2017)
8. Eva, E.O., et al.: Prevalence of stress among medical students: a comparative study between public and private medical schools in bangladesh. BMC. Res. Notes **8**(1), 1–7 (2015)
9. Ge, F., Zhang, D., Wu, L., Mu, H.: Predicting psychological state among chinese undergraduate students in the covid-19 epidemic: a longitudinal study using a machine learning. Neuropsychiatric Disease Treat. **16**, 2111–2118 (2020)
10. Ghrouz, A.K., Noohu, M.M., Dilshad Manzar, M., Warren Spence, D., BaHammam, A.S., Pandi-Perumal, S.R.: Physical activity and sleep quality in relation to mental health among college students. Sleep Breathing **23**, 627–634 (2019)
11. Henry, S.K., Grant, M.M., Cropsey, K.L.: Determining the optimal clinical cutoff on the CES-d for depression in a community corrections sample. J. Affect. Disord. **234**, 270–275 (2018)
12. Jungbluth, C., MacFarlane, I.M., Veach, P.M., LeRoy, B.S.: Why is everyone so anxious?: an exploration of stress and anxiety in genetic counseling graduate students. J. Genet. Couns. **20**(3), 270–286 (2011)
13. Mao, Y., Zhang, N., Liu, J., Zhu, B., He, R., Wang, X.: A systematic review of depression and anxiety in medical students in china. BMC Med. Educ. **19**(1), 1–13 (2019)
14. McGorry, P.D., Killackey, E.J.: Early intervention in psychosis: a new evidence based paradigm. Epidemiology Psychiatric Sci. **11**(4), 237–247 (2002)
15. Moutinho, I.L.D., et al.: Depression, stress and anxiety in medical students: a cross-sectional comparison between students from different semesters. Revista da Associação Médica Brasileira **63**, 21–28 (2017)
16. Womble, M., Jennings, S., Schatz, P., Elbin, R.: A-173 clinical cutoffs on the state-trait anxiety inventory for concussion. Arch. Clin. Neuropsychol. **36**(6), 1228–1228 (2021)

Regarding depression: Concerning study duration, age, academic efficacy scores from the MBI questionnaire, and QCAE affective empathy scores, the corresponding p-values are 0.851, 0.626, 0.578, and 0.405, all significantly greater than 0.05. From a statistical standpoint, this indicates that these features do not manifest significant differences at the given level. In other words, we lack sufficient evidence to support significant relationships or disparities between these features and anxiety.

However, in the context of gender, history of psychological counseling, native language, health status, and MBI Cynicism scores, the corresponding p-values are all below 0.05. This implies that these features may possess some degree of correlation, association, or influence with anxiety. In terms of statistical analysis, these divergences suggest that these features might hold a certain impact or role in relation to anxiety emotions.

5 Conclusion

In this study, we investigated the statistical relationships between various factors and the occurrence of psychological disorders, revealing patterns of variation in the proportions of individuals affected by psychological disorders and the severity of these disorders across different populations. We identified several factors closely associated with psychological disorders, with gender, native language, and health status potentially exhibiting more significant correlations with anxiety and depression.

Nevertheless, our study does have certain limitations. The size of the dataset is relatively small, and the number of features is limited, which could potentially impact the accuracy of our conclusions. To arrive at more universally applicable conclusions, we require a more comprehensive dataset of medical student information and a larger sample size.

Acknowledgment. This work is supported by the Shenzhen Science and Technology Innovation Commission (Stabilisation Support Programme).

References

1. Ahad, A., Chahar, P., Haque, E., Bey, A., Jain, M., Raja, W.: Factors affecting the prevalence of stress, anxiety, and depression in undergraduate indian dental students. J. Educ. Health Promot. **10**, 266 (2021)
2. Al-Dabal, B.K., Koura, M.R., Rasheed, P., Al-Sowielem, L., Makki, S.M.: A comparative study of perceived stress among female medical and non-medical university students in dammam, saudi arabia. Sultan Qaboos Univ. Med. J. **10**(2), 231 (2010)
3. Behere, S.P., Yadav, R., Behere, P.B.: A comparative study of stress among students of medicine, engineering, and nursing. Indian J. Psychol. Med. **33**(2), 145–148 (2011)
4. Carrard, Valerie, C., et al.: The relationship between medical students' empathy, mental health, and burnout: a cross-sectional study. Med. Teacher **44**(12), 1392–1399 (2022)

The presence of a job or a partner appears to have limited influence on anxiety or depression.

4.2 Correlation Analysis

Through t-tests and chi-squared tests, we will determine features that exhibit robust correlations with anxiety and depression, as well as those with weaker correlations.

Table 2. Demographic characteristics of college students - Anxiety and Depression

Variable	Anxiety p-Value	Depression p-Value
Gender	0.000***	0.000***
Job	0.439	0.018**
Part	0.242	0.012**
psyt	0.000***	0.000***
year	0.083*	0.084*
age	0.76	**0.626**
glang	0.000***	0.000***
stud_h	**0.987**	**0.851**
health	0.001***	0.000***
qcae_cog	0.216	0.15
qcae_aff	0.074*	**0.405**
mbi_ex	**0.587**	0.053*
mbi_cy	0.005***	0.000***
mbi_ea	**0.529**	**0.578**

Note: ***, **, * represent significance levels of 1%, 5%, and 10%, respectively.

Based on the results from Table 2, the following insights can be derived:

For anxiety disorder: In the examination of study duration, the significance p-value is 0.987; concerning the emotional exhaustion and academic efficacy scores from the MBI questionnaire, the respective p-values are 0.587 and 0.529; regarding the presence of a job, the p-value is 0.439. These outcomes indicate that statistically, the aforementioned features do not exhibit significant differences at the given level. In other words, we lack sufficient evidence to support a significant association or disparity between these features and anxiety.

However, when considering features such as gender, history of psychological counseling, native language, and health status, all respective p-values are below 0.05. This suggests the potential existence of some degree of correlation, association, or influence between these features and anxiety. This statistical divergence implies that these features might have a certain impact or role in relation to anxiety emotions.

Statistical Analysis of Other Variables' Relationship with Psychological Disorders. The statistical graphs illustrating the proportions of anxiety and depression, as well as the average scores of the affected population, varying with different features, are presented in Fig. 2.

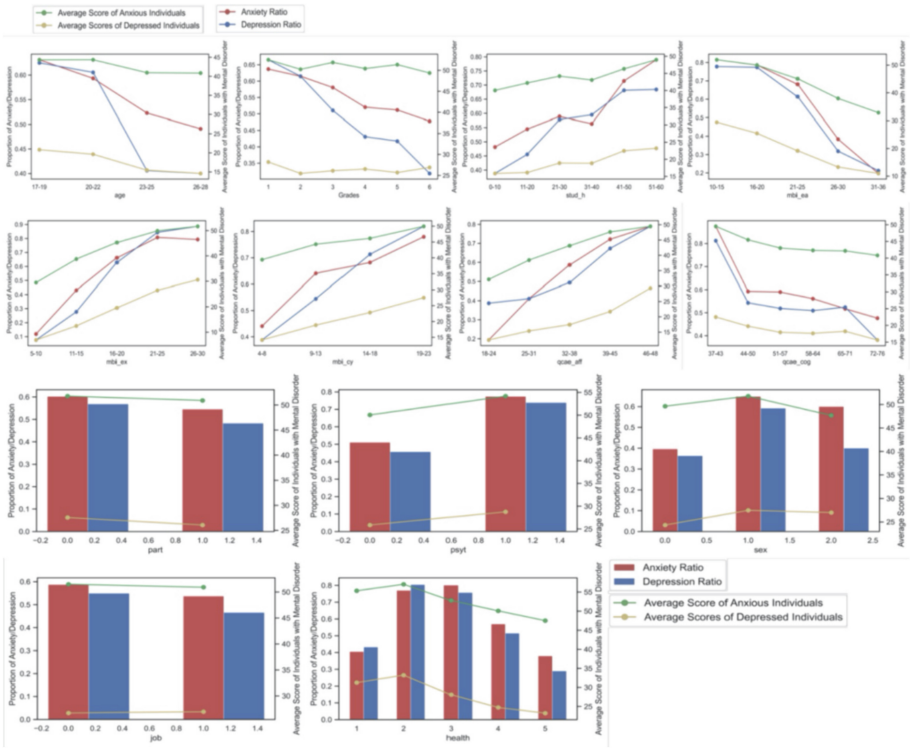


Fig. 2. Anxiety or depression statistical graphs

Features that exhibit a negative correlation with the proportion and severity of individuals with anxiety and depression include: age, academic year, academic efficacy, cognitive empathy, and health status. We observed that as age increases or academic year advances, the proportion of individuals with anxiety or depression decreases. Notably, the proportion of individuals with depression significantly drops after the age of 23. This may be attributed to medical students gradually adapting to the pace of learning, acquiring effective study methods, and consequently reducing the occurrence of anxiety and depression.

Features that show a positive correlation with the proportion and average scores of individuals with anxiety and depression include study duration, emotional exhaustion, cynicism, and affective empathy. We found that individuals with longer study durations exhibit a higher prevalence of psychological disorders, coupled with increased severity.

4 Result and Discussion

4.1 Statistical Analysis

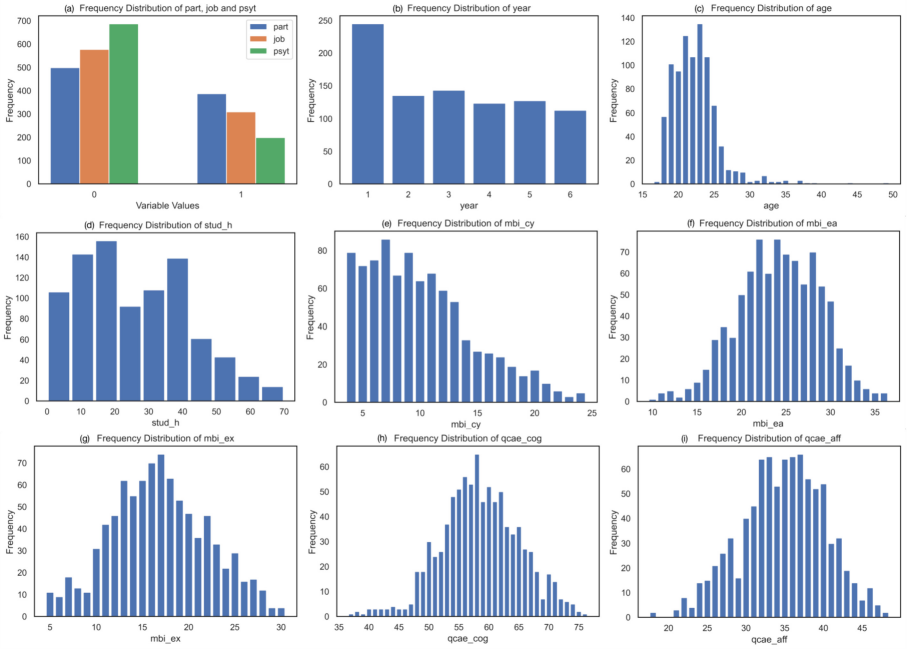


Fig. 1. Frequency Distribution of Each Feature. In Figure (a), the horizontal axis scale of 0 and 1 represents no partner (no job, no psychological treatment) and have a partner (job, psychological treatment) categories, respectively.

Univariate Statistical Analysis. The statistical graphs for each feature are depicted in the Fig. 1.

From Fig. 1(a), it is evident that students without partners outnumber those with partners, and similarly, students without jobs exceed those with jobs. Most students have not undergone psychological therapy over the past year.

Figure 1(b) illustrates that the first-year student count significantly surpasses other academic years, while second to sixth-year students exhibit a more even distribution.

Figure 1(c) indicates that the age distribution of medical students in the sample is concentrated between 18 to 25 years.

Figure 1(d), the highest number of individuals falls within the 10 to 20 h per week study time range. Most individuals do not exceed 40 h of study time per week.

The distributions of other features approximate a normal distribution.

Statistical Description. Creating statistical graphs for individual features provides a more intuitive display of the distribution of each feature's quantity, aiding in gaining a deeper understanding of the overall feature distribution within the sample population.

We have chosen two indicators, the proportion of individuals with psychological disorders and the average scores of the affected population, to depict the quantity and severity of psychological disorders. By visualizing the trends of these two indicators in relation to other features, we can gain a clearer insight into the influence of these features on psychological disorders.

Statistical Inference. This study primarily employs two hypothesis testing methods: the t-test and the chi-squared test, to conduct an analysis of dissimilarities among various features.

The independent samples t-test is utilized to compare differences between categorical and quantitative samples (samples A and B). The main steps are as follows:

1. Hypothesis formulation: The null hypothesis assumes no significant difference between samples A and B, while the alternative hypothesis assumes the presence of a difference.
2. Assumption of sampling distribution: Independent samples A and B are assumed to be approximately normally distributed, satisfying the conditions for t-distribution.
3. Calculation of t-value:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}}$$

4. Calculation of confidence interval for means: Using the computed t-value, along with sample sizes and confidence level, the confidence interval for means is calculated, allowing for statistical inference regarding mean differences.

The Pearson chi-squared test is employed for analyzing differences between two categorical sample variables. The statistical measure used is

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i},$$

which assesses the disparity between theoretical frequencies and observed values.

commonly utilized, each specifically designed for screening anxiety and depression symptoms, respectively.

3 Methods

3.1 Dataset Introduction

The dataset [4,5] for this study was released in 2020 and encompasses information from 886 medical students. The features comprise individual demographic details ('age', 'year', 'sex', 'glang', 'part', and 'job'), educational aspects ('stud_h', 'mbi_cy', and 'mbi_ea'), psychological conditions ('qcae_cog', 'qcae_aff', and 'mbi_ex'), and physical well-being ('health'). Two labels describing the psychological disorder status are 'stai_t' and 'cesd', which are derived from the STAI-T and CESD questionnaires respectively. These questionnaires are widely employed for screening anxiety and depression patients. The introduction of each feature is shown in the Table 1.

Table 1. Study variables

variable name	description
age	age at questionnaire
year	curriculum year
sex	gender
glang	mother tongue
part	having a partner
job	have a paid job
stud_h	how many hours per week spend on study
health	How satisfied are you with your health
psyt	consulted a psychotherapist or a psychiatrist for health
qcae_cog	QCAE Cognitive empathy score
qcae_aff	QCAE Affective empathy score
mbi_ex	MBI Emotional Exhaustion
mbi_cy	MBI Cynicism
mbi_ea	MBI Academic Efficacy

3.2 Statistical Analysis

This study primarily engages in statistical description and hypothesis testing of the dataset, aiming to identify the relationships between psychological disorders (anxiety or depression) and various features.

1 Introduction

The psychological well-being of college students has garnered extensive attention. The university phase, which signifies the transition between academic and social realms [10], marks the initial steps of students venturing into the societal arena. However, due to factors such as uncertainty about the future, substantial academic pressure, challenges in interpersonal relationships, and insufficient self-confidence, college students are susceptible to experiencing psychological health issues such as anxiety and depression [12].

Among various academic disciplines, medical students particularly warrant significant concern as they encounter heightened psychological health challenges [2,3,8]. Their prolonged academic duration, substantial academic pressures, and the weight of future employment prospects create a formidable environment. Moreover, the daily exposure to patients' ailments and suffering brings about negative emotions, thereby increasing the likelihood of psychological health problems.

Despite the plethora of research focusing on factors contributing to psychological disorders, there remains a relative scarcity of investigations concentrating on medical students. Consequently, this study primarily revolves around medical students as a specific sample group, delving into their prevalence and severity of psychological disorders. Concurrently, we aim to discern potential factors contributing to the onset of psychological ailments and analyze the varying degrees of correlation between these factors and psychological disorders.

2 Related Work

Many researchers have initiated investigations into the psychological well-being of medical students. Medical students exhibit higher levels of depression, anxiety, and stress symptoms [7,15]. Such psychological disorders as anxiety and depression can potentially have adverse effects on medical students' personal and professional lives, leading to issues like insomnia and even triggering thoughts of suicide [9].

Mao *et al.* [13] found that the occurrence of depression and anxiety among medical students is influenced by a variety of factors, including individual characteristics, socioeconomic status, and environmental factors such as gender, academic year, family structure, family income, parental educational background, and social support. Additionally, Ahad *et al.* [1] revealed that age, gender, employment status, and accommodation situation are significant factors affecting stress levels among medical students. Notably, female students tend to experience higher stress levels, and those engaged in clinical internships face greater stress compared to pre-internship periods. It's noteworthy that the findings by Moutinho *et al.* [15] emphasize significant variations in the psychological well-being of medical students across different semesters.

Early detection and treatment of mental disorders are crucial for achieving favorable recovery outcomes and reducing the risk of relapse [6,14]. Typically, questionnaire surveys are employed for the early screening of anxiety and depression patients. Among these, the STAI-T [16] and CESD [11] questionnaires are

Analysis of Factors Related to Anxiety and Depression in Medical Students

Zheng Jinfang[1,2,3], Pan Jiachen[1,2,3], Zhang Peiyi[1,2,3], Xiao Yi[2,3], and Wang Wei[2,3,4(✉)]

[1] Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
1120200266@smbu.edu.cn

[2] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
xiaoyi@smbu.edu.cn, ehomewang@ieee.org

[3] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China

[4] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. The psychological well-being of university students, particularly those pursuing medical education, has garnered widespread attention. These students hold the potential to shape the future of societal progress, with medical students shouldering a crucial responsibility for the development of overall community health. However, many medical students are susceptible to psychological disorders such as anxiety and depression due to high levels of stress. While numerous studies have investigated factors contributing to the prevalence of psychological ailments in the general population, there has been a limited focus on analyzing this phenomenon specifically among medical students. This study utilizes a sample of 886 medical students, gathering information regarding their personal backgrounds, academic pursuits, psychological states, and physical health conditions. The aim is to discern which subgroups have a higher prevalence of anxiety or depression. Employing statistical analysis, the relationships between various factors and the occurrence of psychological disorders are examined. Through differential analysis, factors with a stronger correlation to psychological disorders are identified. Notably, factors like study duration and emotional fatigue exhibit a positive association with anxiety and depression, while factors such as academic year and academic efficacy demonstrate a negative correlation. Furthermore, gender and health status exhibit robust correlations with the manifestation of anxiety and depression.

Keywords: Anxiety · depression · medical students · correlative factors

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 52–60, 2024.
https://doi.org/10.1007/978-3-031-65126-7_5

34. Williams, J.B., First, M.: Diagnostic and statistical manual of mental disorders. In: Encyclopedia of Social Work (2013)
35. Wongkoblap, A., Vadillo, M.A., Curcin, V., et al.: Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study. JMIR Mental Health **8**(8), e19824 (2021)
36. Wu, J., Zhou, Z., Wang, Y., Li, Y., Xu, X., Uchida, Y.: Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction. In: 2019 International Conference on Multimodal Interaction, pp. 582–588 (2019)
37. Yoon, J., Kang, C., Kim, S., Han, J.: D-vlog: Multimodal vlog dataset for depression detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 12226–12234 (2022)
38. Zhang, H., et al.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18802–18812 (2022)
39. Zheng, W., Yan, L., Wang, F.Y.: Two birds with one stone: knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. IEEE Trans. Affect. Comput. **14**(4), 2595–2613 (2023)
40. Zhou, L., Liu, Z., Yuan, X., Shangguan, Z., Li, Y., Hu, B.: Caiinet: neural network based on contextual attention and information interaction mechanism for depression detection. Digit. Sig. Process. **137**, 103986 (2023)
41. Zhu, Y., Shang, Y., Shao, Z., Guo, G.: Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Trans. Affect. Comput. **9**(4), 578–584 (2017)
42. Zogan, H., Razzak, I., Wang, X., Jameel, S., Xu, G.: Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web **25**(1), 281–304 (2022)

14. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Lin, C., et al.: Sensemood: depression detection on social media. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 407–411 (2020)
17. Mann, P., Matsushima, E.H., Paes, A.: Detecting depression from social media data as a multiple-instance learning task. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE (2022)
18. de Melo, W.C., Granger, E., Hadid, A.: A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Trans. Affect. Comput. **13**(3), 1581–1592 (2020)
19. de Melo, W.C., Granger, E., Lopez, M.B.: MDN: a deep maximization-differentiation network for spatio-temporal depression detection. IEEE Trans. Affect. Comput. 14(1), 578–590 (2021)
20. Meng, Y., Bridge, J., Addison, C., Wang, M., Merritt, C., Franks, S., Mackey, M., Messenger, S., Sun, R., Fitzmaurice, T., et al.: Bilateral adaptive graph convolutional network on CT based covid-19 diagnosis with uncertainty-aware consensus-assisted multiple instance learning. Med. Image Anal. **84**, 102722 (2023)
21. Mitra, V., et al.: The SRI avec-2014 evaluation system. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 93–101 (2014)
22. Organization, W.H., et al.: Depression and other common mental disorders: global health estimates. World Health Organization, Technical Report (2017)
23. Safa, R., Bayat, P., Moghtader, L.: Automatic detection of depression symptoms in twitter using multimodal analysis. J. Supercomput. **78**(4), 4709–4744 (2022)
24. Saldanha, O.L., et al.: Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. NPJ Precision Oncol. **7**(1), 35 (2023)
25. Salekin, A., Eberle, J.W., Glenn, J.J., Teachman, B.A., Stankovic, J.A.: A weakly supervised learning framework for detecting social anxiety and depression. Proc. ACM Interact. Mobile Wearable Ubiquitous Technol. **2**(2), 1–26 (2018)
26. Shangguan, Z., Liu, Z., Li, G., Chen, Q., Ding, Z., Hu, B.: Dual-stream multiple instance learning for depression detection with facial expression videos. IEEE Trans. Neural Syst. Rehabil. Eng. **31**, 554–563 (2022)
27. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: a survey. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8135–8153 (2022)
28. Sotelo, J.L., Nemeroff, C.B.: Depression as a systemic disease. Personalized Med. Psychiatry **1**, 11–25 (2017)
29. Vahia, V.N.: Diagnostic and statistical manual of mental disorders 5: a quick glance. Indian J. Psychiatry **55**(3), 220 (2013)
30. Valstar, M., et al.: Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2014)
31. Valstar, M., et al.: Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2013)
32. Verhoeven, J.E., Révész, D., Epel, E.S., Lin, J., Wolkowitz, O.M., Penninx, B.W.: Major depressive disorder and accelerated cellular aging: results from a large psychiatric cohort study. Mol. Psychiatry **19**(8), 895–901 (2014)
33. Wang, T., Li, C., Wu, C., Zhao, C., Sun, J., Peng, H., Hu, X., Hu, B.: A gait assessment framework for depression detection using kinect sensors. IEEE Sens. J. **21**(3), 3260–3270 (2020)

detection task, and the best performance results illustrate the effectiveness and superiority of our proposed method. We hope that our work can add more effective contributions to the field of weakly supervised depression detection. In future work, we hope to add more modal social media such as text for depression detection.

Acknowledgment. This work was supported by the National Nature Science Foundationof China (62102266, 62231020, 62272317), Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University, Public Technology Platform of Shenzhen City (GGFW2018021118145859), Shenzhen Science and Technology Innovation Commission (R2020A045), Natural Science Foundation of Guangdong Province-Outstanding Youth-Program (2019B151502018), Pearl River Talent Recruitment Program ofGuangdong Province (2019ZT08X603, 2019JC01X235), National Key R&DProgram of China (2020YFA0908700), and the Natural Science and Engineering Research Council of Canada (corresponding author: Xiping Hu,huxp@bit.edu.cn).

References

1. Al Jazaery, M., Guo, G.: Video-based depression level analysis by encoding deep spatiotemporal features. IEEE Trans. Affect. Comput. **12**(1), 262–268 (2018)
2. Alghowinem, S., Goecke, R., Wagner, M., Parker, G., Breakspear, M.: Eye movement analysis for depression detection. In: 2013 IEEE International Conference on Image Processing, pp. 4220–4224. IEEE (2013)
3. Bourke, C., Douglas, K., Porter, R.: Processing of facial emotion expression in major depression: a review. Aust. NZ. J. Psychiatry **44**(8), 681–696 (2010)
4. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Commun. **71**, 10–49 (2015)
5. Cummins, N., Sethu, V., Epps, J., Schnieder, S., Krajewski, J.: Analysis of acoustic space variability in speech affected by depression. Speech Commun. **75**, 27–49 (2015)
6. Feng, J., Zhou, Z.H.: Deep miml network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
7. Ge, Y., Zhou, Q., Wang, X., Shen, C., Wang, Z., Li, H.: Point-teaching: weakly semi-supervised object detection with point annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 667–675 (2023)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
9. Gui, T., et al.: Cooperative multimodal approach to depression detection in twitter. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 110–117 (2019)
10. Hashimoto, N., et al.: Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3852–3861 (2020)
11. Hendra, C., Pratanwanich, P.N., Wan, Y.K., Goh, W.S., Thiery, A., Göke, J.: Detection of m6a from direct RNA sequencing using a multiple instance learning framework. Nat. Methods **19**(12), 1590–1598 (2022)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433 (2016)

As show in Fig. 1, the max pooling operation performs the worst, but is comparable to machine learning-based methods. The mean pooling operation outperforms the max pooling operation and achieves comparable results to multiple deep network based methods. In contrast, our proposed attention-based pooling operation achieves the best result, which shows that the attention mechanism effectively improves the performance of the MIL framework.

4.5 Effect of Instance Temporal Size

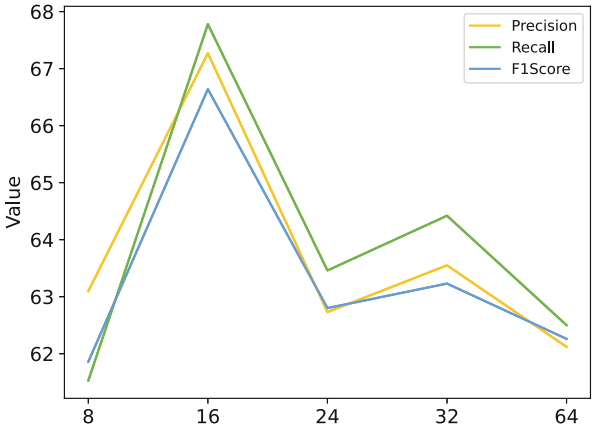


Fig. 2. Evaluation of varying instance temporal size on D-Vlog dataset

To assess the effect of instance time segments size on the method, we construct time segments k of different sizes for the study. As shown in Fig. 2, the best results are achieved in all metrics when k = 16. Moreover, it is worth noting that the results of the model do not exhibit linear change when the value of k increases or decreases, demonstrating that the smaller or larger time segments are not appropriate in depression detection. Technically, the size of the time segment determines the length of the time information contained in the instance. When the size of time segment decreases, continuous time series entering the time window is too short to perform sufficient feature aggregation through the multiple instance pooling layer. When the size of time segment increases, the redundancy of too much temporal information may affect the training of the model.

5 Conclusion

In this study, we perform an attention-based multiple instance learning method to detect depression using social media. We conduct sufficient experiments on the D-Vlog dataset and report the state-of-the-art model performance compared to us. The experimental results show the advantages and potential of multiple instance learning in depression

Table 1. Evaluation of the proposed methods on D-Vlog dataset

Method	Precision(%)	Recall(%)	F1Score(%)
LR	54.86	54.72	54.78
SVM	53.10	55.19	52.97
RF	57.69	58.49	57.84
KNN-Fusion	57.86	59.43	54.25
BiLSTM	60.81	61.79	59.70
TFN	61.39	62.26	61.00
Transformer_Concat	62.51	63.21	61.10
Transformer_Add	59.11	60.38	58.11
Transformer_Multiply	63.48	64.15	63.09
Depression_Detector	65.40	65.57	63.50
Temporal Convolutional	65.40	64.70	65.00
CAIINET	66.57	66.98	66.56
Ours	**67.27**	**67.77**	**66.64**

4.4 Comparison with the MIL Methods

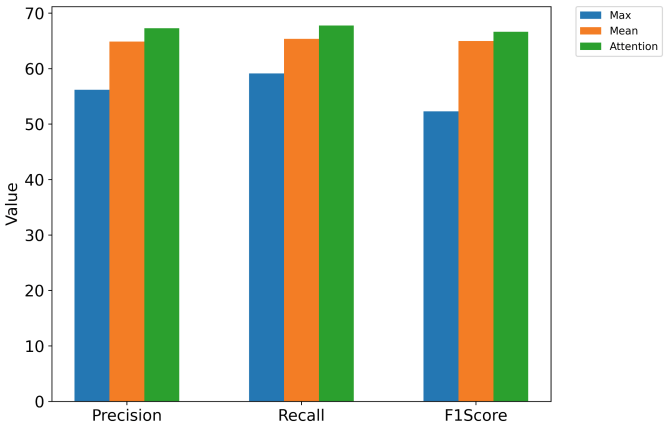


Fig. 1. Evaluation of the MIL methods on D-Vlog dataset

In order to explore the potential of MIL in depression detection and compare with the attention-based MIL method used in this paper, we present the methods based on max pooling [6] and mean pooling [6] as comparison experiments. Notably, the max pooling and mean pooling operation select the feature with the highest and average feature among the instance to obtain the bag-level feature.

4 Experiments

4.1 Experimental Dataset

In this work, we use the D-Vlog dataset [37] collected from YouTube, which contains 961 vlogs videos from 816 subjects composed of 322 males and 639 females. The dataset has a total of 505 depressed subjects and 406 healthy controls, and the average length of the vlog is 596 s. According to the ratio of 7:1:2, the dataset is divided into training set, validation set, and test set, respectively. The preliminary label assignment of the dataset comes from the title keywords of the vlog. Usually, vlogs containing keywords such as "depression daily vlog", "depression journey vlog" and "depression vlog" are labeled as depressed vlog. In addition, vlogs containing keywords such as "daily vlog" and "haul vlog" are labeled as non-depressed vlogs. Then, two tasks are used to ensure the plausibility of labels. First, videos that do not conform to the "vlog" format (e.g., videos without appearance) are removed. Second, specific annotators are assigned to judge whether the subjects have depression by watching the vlog videos with automatic text generation. For privacy protection considerations, D-Vlog only provides the features of the extracted voice and facial expression in the video, which are the 15-dimensional extended Geneva Minimalistic Acoustic Parameter Set and the 68-dimensional facial landmarks, respectively.

4.2 Experimental Setup

In this paper, the size of the time segment that constitutes the instance is 16, and the total length of the bag is limited to 596. All models are trained for 30 epochs, using Adam [15] as the optimizer with learning rate, weight decay and eps are set to 1e-4, 5e-4, and 1e-8, respectively and the batch size is set to 16. We report weighted average precision, recall, and F1 score to evaluate model performance. In order to prevent overfitting, the model uses an early stopping mechanism during training. All experiments are implemented in pytorch, running on a server with NVIDIA 1660 s and 16 GB RAM.

4.3 Comparison with the Previous State-of-the-Art Models

We compare with current state-of-the-art methods to evaluate the effectiveness of our method, and the results are shown in Table 1. Specifically, the recent methods for comparison include: 10 methods using in the D-Vlog dataset [37] as the baseline, the Knowledge-Embedded Temporal Convolutional Transformer method proposed by Zheng et al. [39] and the CAIINET method proposed by Zhou et al. [40]. The traditional machine learning methods including LR, SVM and RF don't perform well on the D-vlog dataset, which is due to the lack of nonlinear fitting ability of machine learning. Moreover, corresponding deep learning methods including BISTM, TFN and Depression Detector achieve better performance compared to machine learning methods.

Compared with baseline, our proposed method improves at least 1.87%, 2.2% and 3.14% on the weighted average precision, recall and F1 score metrics. In addition, compared with the recently proposed Knowledge-Embedded Temporal Convolutional Transformer method and the CAIINET method, our proposed method has achieved the highest results in all metrics, indicating the effectiveness of the proposed method.

$$F_{max} = W_2 h_{max} \tag{6}$$

where W_1 and W_2 represent trainable weights respectively. Moreover, in order to integrate the information of the obtained vectors F_{mean} and F_{max}, we concat them and use a one-dimensional convolution operation to obtain the contextual kernel α:

$$\alpha = f_c\left([F_{mean} : F_{max}]\right) \tag{7}$$

where f_c represents the convolution operation with convolution kernel size is 1. Formally, by combining the context kernel α with the final output O of BILSTM, we obtain the instance features with temporal context. This step can be formulated into:

$$z = \sum_{t=1}^{T} a_t O_{w,t} \tag{8}$$

where,

$$a_t = \frac{\exp\left(\alpha O_{w,t}^{\top}\right)}{\sum_{\tau=1}^{T} \exp\left(\alpha O_{w,\tau}^{\top}\right)} \tag{9}$$

Technically, a_t is the attention weight to indicate the effectiveness of the BiLSTM output feature. Also, instance feature with temporal information is obtained by combining the attention weight with the output feature, which helps to articulate the dynamic information of depressed patients.

3.3 AD-MIL

Recently, many studies have attempted to use attention mechanisms to integrate them into the MIL framework [11,14,38]. Notably, Ilse et al. [14] demonstrate that MIL based on attention pooling can achieve better performance compared to conventional multiple instance pooling such as max pooling and mean pooling. Inspired by these, we use attention pooling to aggregate the instance features obtained in the previous section.

Formally, we denote $Z = \{z_1, \ldots, z_M\}$ as a bag of M instance features, and attention-based MIL pooling can be defined as:

$$e = \sum_{m=1}^{M} b_m z_m \tag{10}$$

with,

$$b_m = \frac{\exp\left\{q^{\top} \tanh\left(V z_m^{\top}\right)\right\}}{\sum_{k=1}^{M} \exp\left\{q^{\top} \tanh\left(V z_k^{\top}\right)\right\}} \tag{11}$$

where q and V are trainable parameters and hyperbolic tangent $\tanh(\cdot)$ is the element-wise non-linearity. In addition, b_m represents as an attention weight indicating the contribution of a given instance to the prediction of the whole bag. Therefore, different attention weights can be used as an implicit feature selection to make the final bag features more informative.

m-th instance of i-th bag. Furthermore, each instance s_{im} is assumed to have implicit label $y_m \in \{0,1\}$ to represent negative or positive, which is not given in practice due to labeling difficulties.

Traditional MIL meets the following constraints: A bag is positive if there is at least one positive instance, while a negative bag is only if all the instances making up the bag are negative. Formally, it follows that

$$Y = \begin{cases} 0, & \text{iff } \sum_m y_m = 0 \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

However, in the case of our work, there will be cases where both positive and negative instances are included in one bag, so assumption here is not strict. Hence, we propose an attention-based algorithm for depression detection by introducing a looser version of the attention mechanism to assign implicit weights to instances.

3.2 AD-LSTM

The proposed AD-LSTM module first uses Bi-directional LSTM (BiLSTM) [8] to extract the temporal information of the two directions of LSTM [12] as output, and then uses the attention mechanism to integrate the semantic features with temporal information to obtain the feature of the instance.

We develope BiLSTM combining information in both directions of LSTM at the same time to obtain richer semantic information in the instance. Notably, each layer of BiLSTM consists of LSTM in two directions, and the outputs of the layer are as follows:

$$h_{i,t} = l_f\left(O_{i-1,t}\right) \tag{2}$$

$$H_{i,T-t} = l_b\left(O_{i-1,T-t}\right) \tag{3}$$

$$O_{i,t} = \left[h_{i,t}, H_{i,T-p}\right] \tag{4}$$

where T represents the total length of the segment. l_f and l_b represent the forward and backward LSTM models, respectively. $h_{i,t}$ and $H_{i,T-t}$ represent the output of the i-th layer at the time t of the forward LSTM and the output of the i-th layer at the $T-t$ time of the backward LSTM. Then, we add the forward and backward features of the last layer w of BILSTM to get $O = \{O_{w,1}, \ldots, O_{w,T}\}$, and we connect the forward and backward output features of BILSTM at time T to form contextual feature $h = \{h_{1,T}, H_{1,T}, \ldots, h_{w,T}, H_{w,T}\}$. Similar to the feature map in CNN, each $h_{i,T}$ in h represents the feature of BILSTM at the last time. Therefore, in order to obtain rich contextual information, we use the mean pooling operation and max pooling operation, which is commonly used in spatial information processing, to obtain context features h_{mean} and h_{max}. Further, we use the two-layer network model to introduce the non-linear operations of the two pooling features:

$$F_{mean} = W_1 h_{mean} \tag{5}$$

2.2 Weak Supervision and Multiple Instance Learning

Multiple instance learning (MIL) is a form of weakly supervised learning, which is used to deal with model training under insufficient labels. Typically, in multiple instance learning, the model only receives coarse-grained bag-level labels, and the labels of the instances that make up the bag are unknown. According to the different MIL settings, the current MIL algorithm can be divided into instance-level [13] and bag-level [10]. Due to the difficulties and high costs in the actual labeling process, specific annotators can only assign bag-level labels in the context. Hence, MIL has been widely used in many fields including object detection [7], pneumonia detection [20] and tumor detection [24].

Recently, several works of MIL has been applied for depression detection. Concretely, in the use of weakly supervised learning framework, Salekin et al. [25] proposed a MIL method to identify depression from voice speech containing labels of depressed patients without providing specific segments of symptoms. Shangguan et al. [26] proposed a dual-stream MIL deep network to identify depression by using raw facial expressions. In addition, extensive works have used MIL for detecting depression on social media due to its superiority. Wongkoblap et al. [35] proposed two multiple instance learning models to predict depression using textual information from Twitter. Moreover, a MIL method for detecting depression using students' posts from university was presented by Mann et al. [17]. They performed theoretical and experimental analysis by using Transformer and LSTM model on the dataset of university students.

Previous work using MIL to identify depression in social media has mainly focused on text information, and few works have used the information of subjects' facial expressions and voices to identify depression. Since the facial expressions and voice can express the mental state of the subjects and the effectiveness of MIL in the detection of depression, it is very necessary to establish a model for detecting depression using MIL based on these two modalities. Inspired by the work of these pioneers, we aim to expand the scope of the current literature on depression detection through the application of MIL and attention mechanisms.

3 Methods

We propose a weakly supervised learning model for the depression detection task in a single end-to-end deep network. Concretely, our model receives the vocal features and visual features extracted by OpenSmile and Dlib respectively, and then the AD-LSTM module extracts the temporal information within the instances. Finally, the AD-MIL module integrates the information of the instance for identifying depression. In this section we present the formulation of MIL and the details of the proposed MIL model.

3.1 Preliminaries

The MIL algorithm receives N labeled sample pairs $\mathrm{D} = \{(S_1, Y_1), \ldots, (S_N, Y_N)\}$, where S_i (i from 1 to N) is the whole bag and Y_i is $\{0, 1\}$ for binary classification of depression and health. Also, $S_i = \{s_{i1}, s_{i2}, \ldots, s_{iM}\}$ consists of M instances where s_{im} represents the

approaches using weakly supervised learning. In Sect. 3, we introduce the relevant preliminaries and the details of our proposed model. We provide the datasets, experimental settings and results used in the experiments in Sect. 4. Finally, we conclude our work in Sect. 5.

2 Related Work

This section briefly reviews the related methods of depression detection and weakly supervised learning.

2.1 Deep Learning for Depression Detection

Since the emerging applications in affective computing, the deep learning-based methods can use behavior signals for depression detection. The datasets for depression detection tasks can be divided into task-specific collection and non-specific task collection. In a specific task, the process of data collection comes from recording subjects completing a certain task according to the requirements of the examiner, such as answering some specific questions or discussing the certain topics. In a non-specific task, the process of data collection comes from external information, such as, voice, video, and text of individuals on the Internet.

AVEC2013 [31] and AVEC2014 [30] are typical task-specific datasets which focus on video modalities. For example, Zhu et al. [41] proposed a two-stream deep network to detect depression by considering the appearance and movements of subjects. By leveraging the optical flow of dynamic information of facial expressions, they improved the performance of the model. Similarly, Jazaery et al. [1] used a convolutional 3D network (C3D) to capture spatio-temporal information and to learn the features of continuous segments through Recurrent Neural Network (RNN). To reduce the model size for depression detection, Melo et al. [19] proposed the 2D deep network (a.k.a., MDN) to capture the spatio-temporal information in facial videos. By embedding the maximization block and difference block in the 2D deep network, the model captured the subtle changes and sudden transitions between face expression, and achieved comparable performance to 3D deep network.

Since a considerable number of users share recent life emotions and states on the Internet, social media can provide data information under non-specific tasks for depression detection. There are several approaches that use multi-modal data of social media for depression detection. For example, Safa et al. [23] used the biological features, features generated by analyzing user profile pictures, and banner images to detect depression. By using image and text information posted by users on social media, Gui et al. [9] introduced a new collaborative multi-agent reinforcement learning method to predict depression. Zagan et al. [42] presented a novel interpretable depression detection framework, the Hierarchical Attention Network, which used textual, behavioral, temporal, and semantic aspects of social media features for deep learning. Moreover, a deep visual-textual multimodal learning system dubbed SenseMood was proposed to predict the mental state of the users on social networks. Lin *et al.* [16] used CNN and Bert to extract deep representations of pictures and text on social media, which were combined for further depression classification.

of depression mainly relies on the subjective and complex reports of the subjects and the professional judgment of the psychiatrists. For example, the clinician rating scales (e.g., Hamilton rating scale) require rigorous training of raters. The self-rating depression scales (e.g., Self-rating depression scale) rely on accurate description, assessment, and expression of subjects and may change the purpose of their report [29]. Due to the lack of medical resources, the great harm of depression, and the large number of patients, the subjective assessment and diagnosis cannot meet the current demands for depression diagnosis. In this vein, the automatic detection of depression has attracted ever-increasing research attention due to objectivity, fast deployment, and long endurance.

With the advancement of affective computing, previous studies use behavioral signals as objective indicators to conduct research on depression detection, which provides an objective and effective way for auxiliary diagnosis of depression. Many current research outcomes have shown that common behavioral signals can be used as objective indicators for depression detection, such as, eye movement [2], voice [4], gait [33], and facial expression [3].

Different from eye movements and gaits that need to be collected during professional experiments, the leveraged facial expressions and voices in our research obtained through more relaxed methodologies (e.g., social media). In this paper, we use social media data collected from vlog of people documenting their daily lives on the Internet. Compared with data collected from the experimental environments, vlog data has three advantages: 1) easier to obtain; 2) larger quantity; and 3) consistent increase in volume. The three advantages of the vlog data allow to build a more generalized model and explore the ability to articulate the datasets in the wild. In general, a vlog dataset has both facial expressions and voice modalities, and there are dedicated annotators to judge whether the subjects are depressed. However, a complete piece of vlog data has only one binary classification sparse label, because it is impractical for annotators to accurately label the symptoms reflecting depression at a fine-grained level. The traditional depression detection methods [1,5,18,21] assign the same coarse-grained labels to the training instances (video clips or single frames) and may lead to overfitting and corresponding performance loss [27,36].

To address this challenge, we propose a weakly supervised method to identify the binary classification of the subjects (depression or health) using vlog data. Our model takes temporal segments of a certain size as instance input, and uses AD-LSTM to extract temporal contextual information to obtain instance-wise representations. Then, AD-MIL views the vlog video of each subject as a bag containing multiple instances that may be positive or negative. More specifically, the AD-MIL model first uses an attention mechanism to identify the contribution weights of instances to the final classification, and then obtains individual subject representations by combining the weights with instance representations. Technically, using the attention mechanism can effectively alleviate the impact of instances that are not related to the classification label and integrate the information of the whole bag. We conduct a series of experiments on the D-vlog dataset [37], and the experimental results show that our proposed method exceeds the state-of-the-art works, indicating the effectiveness of the proposed method.

The remaining parts of this work are organized as follows. In Sect. 2, we present recent approaches to automatic depression detection using deep learning as well as

Automatic Depression Detection Using Attention-Based Deep Multiple Instance Learning

Zixuan Shangguan[1], Xiaxi Li[2], Yanjie Dong[2], and Xiaoyan Yuan[1(✉)]

[1] Beijing Institute of Technology, Beijing, China
yxy_newly@163.com
[2] Shenzhen MSU-BIT University, Shenzhen, China
{1120200239,ydong}@smbu.edu.cn

Abstract. Depression is a serious mental illness and one of the leading causes of suicide worldwide. However, the social prejudice and the lack of psychiatrists for depression lead to a significant number of depressed patients without accurate diagnosis and subsequent serious consequences. With the rise of social media, previous studies have found that the information of depressed patients on social media can be analyzed to automatically detect depression for auxiliary diagnosis. In the context of weakly supervised learning framework, a multiple instance learning (MIL) method is proposed to identify depression from social media with visual and vocal information. By leveraging the state-of-the-art attention-based deep LSTM (AD-LSTM), the proposed MIL method can handle the problem with sparse labels (i.e., one label for a long-term sequence of visual information). More specifically, the AD-LSTM module is used to process a fixed-length visual and vocal segments to extract temporal representations of instances, and the AD-MIL module is used to aggregate the obtained temporal representations for individual subject predictions. Compared with current benchmarks, our experiments demonstrate that our proposed MIL method can achieve the best weighted average precision, recall and F1 score with the corresponding values as 66.56%, 66.98% and 66.55%, respectively. The numerical results illustrate that the potential and effectiveness of our proposed MIL method in the field of depression detection.

Keywords: Depression Detection · Multiple Instance Learning · Social media

1 Introduction

Major depressive disorder (a.k.a.. depression) is a critical mental illness with serious consequences for individual physical and mental health. More than 300 million people worldwide, which is equivalent to 4.4% of the global population, are currently suffering from varying degrees of depression [22]. Depressed people often exhibit low mood, loss of interest in practice, sleep disturbance, loss of appetite, lack of self-confidence, loss of energy, and inability to concentrate [34]. In addition, depression increases the risk of diabetes, heart disease, Alzheimer's and, in more severe cases, suicide [28,32].

Accurate diagnosis of depression can be effectively controlled and treated through psychological consulting and psychotropic medication. However, the current diagnosis

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 40–51, 2024.
https://doi.org/10.1007/978-3-031-65126-7_4

29. Shen, G., et al.: Depression detection via harvesting social media: a multimodal dictionary learning solution. In: IJCAI, pp. 3838–3844 (2017)
30. Shen, G., e al.: Depression detection via harvesting social media: a multimodal dictionary learning solution. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 3838-3844. AAAI Press (2017)
31. Tadesse, M.M., Lin, H., Xu, B., Yang, L.: Detection of suicide ideation in social media forums using deep learning. Algorithms **13**(1), 7 (2019)
32. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences (2018)
33. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans. Knowl. Data Eng. **32**(3), 588–601 (2018)
34. Wang, N., etal.: Learning models for suicide prediction from social media posts. arXiv preprint arXiv:2105.03315 (2021)
35. Wang, Y., Wang, Z., Li, C., Zhang, Y., Wang, H.: A multitask deep learning approach for user depression detection on sina weibo. arXiv preprint arXiv:2008.11708 (2020)
36. Yang, T., et al.: Fine-grained depression analysis based on chinese micro-blog reviews. Inf. Process. Manage. **58**(6), 102681 (2021)
37. Yao, X., Yu, G., Tang, J., Zhang, J.: Extracting depressive symptoms and their associations from an online depression community. Comput. Hum. Behav. **120**, 106734 (2021)
38. Zhou, S., Zhao, Y., Bian, J., Haynos, A.F., Zhang, R., et al.: Exploring eating disorder topics on twitter: machine learning approach. JMIR Med. Inform. **8**(10), e18273 (2020)
39. Zogan, H., Razzak, I., Jameel, S., Xu, G.: Depressionnet: a novel summarization boosted deep framework for depression detection on social media. ArXiv **abs/2105.10878** (2021)

13. Gui, T., et al.: Cooperative multimodal approach to depression detection in twitter. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19, AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.3301110
14. Holt-Lunstad, J., Smith, T.B., Baker, M., Harris, T., Stephenson, D.: Loneliness and social isolation as risk factors for mortality: a meta-analytic review. Perspect. Psychol. Sci. J. Assoc. Psychol. Sci. **10**(2), 227 (2015)
15. Kessler, R.C., et al.: Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative. World Psychiatry **6**(3), 168 (2007)
16. Kohler, C.G., Hoffman, L.J., Eastman, L.B., Healey, K., Moberg, P.J.: Facial emotion perception in depression and bipolar disorder: a quantitative review. Psychiatry Res. **188**(3), 303–309 (2011)
17. Lewis, M., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.703, null ; Conference date: 05-07-2020 Through 10-07-2020
18. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced NLP tasks. arXiv preprint arXiv:1911.02855 (2019)
19. Lin, H., Jia, J., Nie, L., Shen, G., Chua, T.S.: What does social media say about your stress? In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, pp. 3775-3781. AAAI Press (2016)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
21. Malhi, G.S., Mann, J.J.: Depression. The Lancet **392** (2019)
22. Malhotra, A., Jindal, R.: Deep learning techniques for suicide and depression detection from online social media: a scoping review. Appl. Soft Comput. **130**, 109713 (2022)
23. Meng, Y., Li, M., Li, X., Wu, W., Li, J.: Dsreg: using distant supervision as a regularizer. arXiv preprint arXiv:1905.11658 (2019)
24. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in twitter. In: Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012, pp. 1–8 (2012)
25. Pyszczynski, T., Holt, K., Greenberg, J.: Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. J. Pers. Soc. Psychol. **52**(5), 994 (1987)
26. Ríssola, E.A., Losada, D.E., Crestani, F.: A survey of computational methods for online mental state assessment on social media. ACM Trans. Comput. Healthcare **2**(2), 1–31 (2021)
27. Salas-Zárate, R., Alor-Hernández, G., Salas-Zárate, M.D.P., Paredes-Valverde, M.A., Bustos-López, M., Sánchez-Cervantes, J.L.: Detecting depression signs on social media: a systematic literature review. In: Healthcare, vol. 10, p. 291. MDPI (2022)
28. Sekulić, I., Strube, M.: Adapting deep learning methods for mental health prediction on social media. arXiv preprint arXiv:2003.07634 (2020)

5 Conclusions

In this work, we have devised a versatile framework for processing Twitter data to facilitate the diagnosis of early-stage depression. Through an extensive study, we have ascertained that Twitter textual content, user profile information, and historical posting data all hold profound significance in diagnosing depression. Consequently, we have proposed a comprehensive model that amalgamates these inputs and conducted empirical validations to evince the efficacy of our approach. Moreover, we addressed the issue of imbalanced data pertaining to depression patients by exploring several diverse methodologies, culminating in commendable achievements.

Acknowledgement. This work is supported by the Natural Science Foundation of Guangdong Province of China (No. 2021A1515011905)

References

1. Ahmed, U., Mukhiya, S.K., Srivastava, G., Lamo, Y., Lin, J.C.W.: Attention-based deep entropy active learning using lexical algorithm for mental health treatment. Front. Psychol. **12**, 642347 (2021)
2. Beck, A.T.: Cognitive Therapy of Depression. Guilford Press, New York (1979)
3. Belmaker, R.H., Agam, G.: Major depressive disorder. New England J. Med. Mech. Disease **385**, 47–60 (2008)
4. Birmaher, B., Ryan, N.D., Williamson, D.E., Brent, D.A., Kaufman, J.: Childhood and adolescent depression: a review of the past 10 years. part ii. J. Am. Acad. Child Adolescent Psychiatry **35**(11), 1427–1439 (1996)
5. Brent, A.D.: Course and outcome of child and adolescent major depressive disorder. Child Adolescent Psych. Clin. North Am. **11**(3), 619–637 (2002)
6. Carlson, G.A.: The challenge of diagnosing depression in childhood and adolescence. J. Affect. Disord. **61**(supp-S1), S3–S8 (2000)
7. Castillo-Sánchez, G., Marques, G., Dorronzoro, E., Rivera-Romero, O., Franco-Martín, M., De la Torre-Díez, I.: Suicide risk assessment using machine learning and social networks: a scoping review. J. Med. Syst. **44**(12), 205 (2020)
8. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
9. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
11. Fine, A., Crutchley, P., Blase, J., Carroll, J., Coppersmith, G.: Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data. In: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pp. 50–54 (2020)
12. Ghosh, S., Anwar, T.: Depression intensity estimation via social media: a deep learning approach. IEEE Trans. Comput. Soc. Syst. **8**(6), 1465–1474 (2021)

Table 3. The impact of different information

User information	Prec	Rec	F1	Acc
Historical tweet	80.2%	**89.0**%	**84.3**%	83.5%
Historical tweet+BiGRU	**85.8**%	82.8%	84.2%	**84.5**%

the model's robustness. In practical scenarios, this inclusion can mitigate the risk of misidentifying healthy individuals as depressed patients. Furthermore, as user information tends to display greater individuality and lacks a discernible pattern, our user information extraction module can still capture its distinctive features, thereby contributing to incremental improvements in the overall model.

4.4 The Experiment on Data Imbalance

To investigate the efficacy of traditional methods for handling imbalanced data in depression detection, we randomly extracted two imbalanced datasets from the original dataset. The first dataset had a ratio of depressed patients to non-depressed patients of 1:4, while the second dataset had a ratio of 1:2. Subsequently, we conducted a comparative analysis of three methods on these two datasets: the direct usage of cross-entropy loss without any imbalance treatment, the employment of Focal Loss for handling imbalance, and the utilization of Dice Loss for the same purpose. The experimental outcomes are presented in Table 4, with the evaluation metric being the F1 score.

Table 4. The effects of different methods for handling data imbalance.

Imbalanced ratio	CE	Focal Loss	Dice loss
1:4	33.3%	**79.5**%	75.7%
1:2	66.3%	79.8%	**81.8**%

From Table 4, it can be observed that when the imbalance ratio reaches 1:4, the model fails to learn useful classification knowledge from the limited positive samples when using cross-entropy loss without imbalance handling. Even under a 1:2 imbalance ratio, the results are significantly unsatisfactory. Focal Loss, in comparison to cross-entropy, demonstrates a 46.2% and 13.5% increase in F1 scores at 1:4 and 1:2 imbalance ratios, respectively. Similarly, Dice Loss shows a 42.4% and 15.5% improvement in F1 scores at 1:4 and 1:2 imbalance ratios. This verifies the effectiveness of traditional data imbalance handling methods for depression classification in our model. Additionally, it is evident that under higher imbalance conditions (1:4), Focal Loss outperforms, achieving an F1 score of 79.5%, whereas when the imbalance ratio reduces to 1:2, the performance of Dice Loss is superior, achieving an F1 score of 81.8%.

Historical Tweet Model. We employed the large-scale language pretraining model BERT and attention-based bidirectional LSTM to construct a historical tweet feature extraction model. Subsequently, we conducted experimental comparisons on each module, and the results are presented in Table 2. Note: due to better performance with straightforward extraction during data processing, all experiments were performed on data obtained through straightforward extraction.

Table 2. Historical tweet model

model	Prec	Rec	F1	Acc
BERT	77.7%	82.8%	80.1%	79.5%
BERT+BiLstm	**81.6**%	83.3%	82.4%	82.3%
BERT+StackedBiLstm	80.2%	**89.0**%	**84.3**%	**83.5**%

According to Table 2, it can be observed that the utilization of BERT in conjunction with the StackedBiLSTM model yields the most favorable results when processing textual features. Following this, the employment of BERT in combination with BiLSTM ranks second in performance. We believe this is due to a certain temporal correlation in the user's tweet data. Due to BiLSTM's capacity to maintain "memory," the model with an added BiLSTM layer outperforms the classification model solely relying on BERT. Furthermore, it is evident that the StackedBiLSTM model outperforms the BiLSTM model in terms of recall, F1 score, and accuracy, surpassing it by 5.7%, 1.9%, and 1.2%, respectively. However, it should be noted that the accuracy is reduced by 1.4% . We posit that this could be attributed to an excessive focus on contextual information, leading to the inadvertent capture of some depression-irrelevant data and thereby increasing the likelihood of misidentifying depression patients.

The Impact of User Information. We investigated the impact of user information on the classification of users with depression in our experiment. We extracted personal information from users, including the number of individuals they follow, the count of their followers, the quantity of friends, and the number of tweets sent in the past month. For feature extraction, we utilized a Bidirectional Gated Recurrent Unit (BiGRU) to complement the historical tweet features. We explored the experimental outcomes achieved by solely using historical tweets and by amalgamating user information with historical tweets. The model employed in this study is depicted in the aforementioned model diagram in Fig. 1, and the results are presented in Table 3.

From Table 3, it can be observed that the incorporation of user information and historical text enhances the model's precision and accuracy, surpassing the historical tweet model by 4.6% and 1%, respectively. However, the recall rate decreased by 6.2%. We contend that the inclusion of user information enhances

4.2 Hyperparameter Configuration

We attempted to employ the BERT pre-trained model as our Encoder model. For the tweet extraction model (a BiLSTM classification model with attention mechanism), we utilized a 2-layer BiLSTM with hidden layer neural units set to 128. As for the user behavior model, we opted for a single-layer BiGRU with hidden layer neural units set to 128. All experiments were conducted on an RTX3090GPU using the Pytorch framework. We employed the SGD optimizer for training with specific parameters: learning rate (lr) = 0.001, momentum = 0.9, and weight decay = 0.0004. Additionally, we employed a warm-up strategy to reach the initial learning rate. We performed a total of 30 epochs, and during the 10th and 20th epochs, we applied a learning rate decay with a rate of 0.1.

We introduced two approaches to process tweets: simple tweet extraction and K-means-based tweet filtering. To assess the performance of our model, we employed metrics such as accuracy, precision, recall, and F1 score. Additionally, to examine the impact of various components in the model, we conducted an ablation analysis. In experiments involving imbalanced data, we set $\gamma = 2$ in the Focal Loss and $\alpha = 0.9$ in the Dice Loss, with $\epsilon = 1e^{-4}$.

4.3 Ablation Study

Method of Data Processing. We have presented three distinct approaches for data processing: a straightforward tweet extraction method and a tweet filtering based on K-means clustering. These methods were subjected to experimental comparison. All models employed the BERT pre-trained model in conjunction with a single layer of Bidirectional LSTM (BiLSTM). The results are presented in Table 1.

Table 1. Module for Data Processing

Data processing	Prec	Rec	F1	Acc
simplistic tweet extraction	**81.6**%	83.3%	**82.4**%	**82.3**%
K-means	76.6%	**83.5**%	79.9%	79.0%

As indicated in Table 1, the employment of a simplistic tweet extraction approach exhibits superior performance compared to the utilization of the K-means extraction approach, yielding improvements of 5%, 2.5%, and 3.3% in accuracy, F1 score, and precision, respectively, over the BiLSTM model. In terms of recall, the difference between the two methods is negligible. We hypothesize that this discrepancy could be attributed to K-means clustering, which identifies text closely related to the classification but, at the same time, disrupts contextual coherence to some degree, leading to the deterioration of results.

the model to pay less attention to them. From a derivative perspective, once the model correctly classifies the current sample (just passed the 0.5 threshold), Dice Loss leads the model to pay less attention to it, unlike cross-entropy, which encourages the model to approach the two endpoints, 0 or 1. This effectively prevents the training of the model from being dominated by numerous straightforward samples.

3.4 User Information Model

After considering user behavioral data, we extracted relevant features pertaining to their social interactions, such as the number of followers and friends. Furthermore, we took into account user-generated actions, including the quantity of tweets posted and tweets favorited. These extracted features were utilized as inputs for the Bidirectional Gated Recurrent Unit (BiGRU) [9].

Both GRU and LSTM employ gating mechanisms to capture interdependencies among inputs, with GRU being a simplified variant of LSTM. Given the relatively straightforward nature of user behavioral data, we posit that the Bidirectional GRU is better suited for capturing relationships among these features.

Subsequently, the features derived from the Bidirectional GRU were fed into a fully connected layer. The resulting output from this layer serves as a guiding factor for classifying users within the historical posting model.

4 Experimental Setup

4.1 Dataset

We have reprocessed the extensive publicly available depression dataset proposed by [29]. These tweets were collected and labeled by the authors on Twitter, while also retrieving the user's historical tweets within a month. The dataset consists of three parts: (1) **Depression Patient Dataset D1**, comprising 2506 labeled samples of depressed users and their tweets; (2) **Non-depression Patient Dataset D2**, comprising 4166 labeled samples of non-depressed users and their tweets; and (3) **Depression Patient Candidate Dataset D3**. The author constructed a large-scale unlabeled depression candidate dataset containing 58,810 samples. In our experiments, we only utilized the labeled datasets: D1 and D2. We preprocessed the datasets by removing users with fewer than ten posting histories, users without an anchor tweet, or users posting tweets in languages other than English. Additionally, we removed emojis from the data to eliminate any impact on the experimental results, thus ensuring that we have sufficient statistical information related to each user. Finally, for balanced data experiments, we considered only 4000 user samples, with 2000 samples each for depressed and non-depressed users. For unbalanced data experiments, we explored the ratios of depressed users to non-depressed users at 1:2 and 1:4, with 2000 samples for non-depressed users in both cases. For testing purposes, we randomly divided the dataset into a training set (80%) and a test set (20%).

The mechanism of attention allows assigning distinct weights to each input feature and reflects the correlation between features and outcomes.

Data Imbalance. The phenomenon of data imbalance is quite common within social media datasets. This imbalance gives rise to the following two issues:

(1) Disparity between training and testing procedures: Under the influence of imbalanced data, models tend to converge towards points that strongly favor classes with the majority labels. This, in effect, creates a disparity between the training and testing processes. During training, each training instance contributes equally to the objective function, whereas during testing, F1 equally weighs the contributions of positive and negative samples.
(2) Excessive impact of simple negative samples on the model: As pointed out by [23], an abundance of negative samples implies a large quantity of straightforward negatives. Consequently, an overwhelming proportion of the loss stems from these numerous simple negative samples, thereby dominating the gradients and hindering the model from adequately learning how to differentiate between positive samples and challenging negative samples. Both cross-entropy (CE) and maximum likelihood estimation (MLE), which are extensively utilized loss functions in machine learning, fail to address these two issues.

Focal Loss and Dice Loss are two deliberately designed loss functions aimed at mitigating the imbalance between positive and negative samples during the one-stage classification process.

The primary objective of Focal Loss is to diminish the weight of easy samples, thereby focusing the training on the negation of difficult samples. To be more precise, Focal Loss introduces a modulation factor $(1-p_t)^\gamma$ into the cross-entropy loss, where $\gamma \geq 0$ represents an adjustable focal parameter. The general form of Focal Loss can be expressed as follows:

$$FL(p_t) = -(1-p_t)^\gamma log(p_t). \tag{4}$$

Dice Loss, on the other hand, contemplates the classification task from a distinct perspective. In this framework, categorizing a sample as negative is contingent solely upon its probability being less than 0.5; there is absolutely no need to drive it towards 0. Furthermore, considering that the primary objective is to mitigate the data imbalance issue within the dataset and, consequently, enhance the effectiveness of the F1 evaluation metric, Dice Loss is designed to exert a direct impact on F1.

Consequently, the general formulation of Dice Loss has been derived as follows:

$$Dice(p_t) = \frac{2(1-p_t)p_t \cdot y_t + \epsilon}{(1-p_t)p_t + y_t + \epsilon}, \tag{5}$$

where p_t represents the estimated probability, incorporated ϵ acts as a smoothing term, and y_t denotes the true label. The term $(1-p_t)$ serves as a scaling factor. For uncomplicated samples (when p_t approaches 1 or 0), $(1-p_t)p_t$ prompts

of tweets articulating their emotions over several days. Considering the token length constraints inherent in BERT [10] and the acknowledged temporal impact on a user's narrative, we judiciously adopt a straightforward strategy: selecting a user's most recent 20 tweets as the primary method for tweet processing. In cases where a user has fewer than 20 tweets, we employ padding techniques to ensure completeness of the dataset.

Tweet Filtering Based on K-Means Clustering. We acknowledge that not all of a user's posts are necessarily relevant to the point we focus on. For instance, a depressed individual might also publish tweets expressing positive emotions, such as 'Today's weather is lovely!' In order to mitigate the influence of such tweets on our determination of a user's depressive status, we endeavor to filter out tweets that more accurately portray the user's identity. Since we cannot introduce user labels during the processing phase, we adopt an unsupervised approach to analyze users' historical posts. Consequently, we employ the K-means clustering method as our second approach to tweet processing. We select a user's most recent 50 posts, tokenize them using BERT, apply the K-means algorithm to cluster the tweets into two categories, and then extract the 20 tweets closest to the cluster centroids. Should there be an insufficient number of tweets remaining, we will once again utilize padding to complete the dataset.

3.3 Historical Posting Information Model

We employed a pre-trained BERT model and a bidirectional LSTM (BiLSTM) based on an attention mechanism to process the input, capturing sequential information such as sentence context. Moreover, in light of the minority representation of depression patients, we tackled the prevalent issue of imbalanced data in the depression dataset by adopting the Focal Loss and Dice Loss techniques, as introduced in the works by [18,20], respectively.

Classification Module Based on Pre-trained Bidirectional LSTM. From the embedding layer of BERT, the extracted features are passed to the Bidirectional Long Short-Term Memory (BiLSTM), which is an RNN designed to capture sequential information and the long-term dependencies within sentences. Comprising the Bidirectional LSTM are the forward and backward LSTMs, each one independently updating the input x_i at time t:

$$forward(h_t) = LSTM(x_t, forward(h_{t-1})). \tag{1}$$

$$forward(h_t) = LSTM(x_t, backward(h_{t-1})). \tag{2}$$

After BiLSTM processing, the hidden state h_t at time t is a concatenation of the states $\overrightarrow{h}$ and $\overleftarrow{h}$ obtained from the forward LSTM and backward LSTM, respectively. The representation of the i-th word is as follows:

$$h_t = forward(h_{t-1}). \oplus backward(h_{t-1}). \tag{3}$$

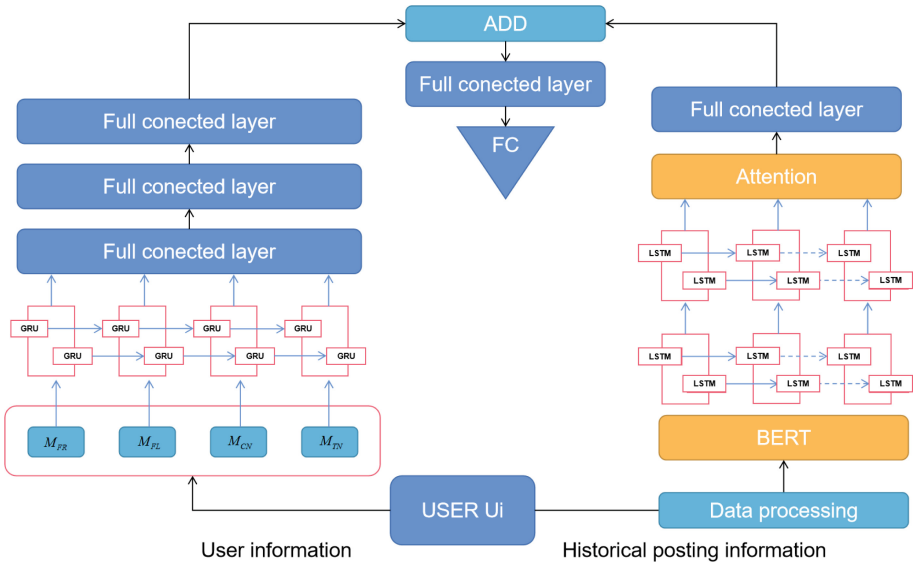


Fig. 1. We have integrated the user information and the user's past tweets of each user to forecast their labels. The user's historical tweets are processed through a tweet processing procedure and then handled by a classification model based on pre-training and bidirectional LSTM. The user's information, on the other hand, is processed using a model based on bidirectional GRU.

extraction, tweet filtering based on K-means clustering. Furthermore, we also incorporate user behavioral metrics, including the number of friends, followers, favorited contents, and the frequency of posts within a month, denoted as $M_{FR}, M_{FL}, M_{CN}, M_{TN}$ respectively.

3.2 Data Processing

The analysis of users' historical tweets constitutes a pivotal aspect of the depression assessment process. Consequently, our efforts are directed towards scrutinizing the past tweets of individuals experiencing depression, in comparison to those of mentally healthy users. This endeavor seeks to afford us a more profound understanding of the behavioral patterns manifested by individuals grappling with depression. Following this analysis, our objective is to leverage these insights to enhance the efficacy of depression detection methodologies.

Conducting Direct Tweet Extraction. Primarily, our initial consideration revolves around a streamlined approach to tweet handling, specifically involving the extraction of a defined quantity of tweets to encapsulate the entirety of a user's Twitter activity. We posit that the temporal dynamics of a user's posts exert a substantial influence on the thematic content of their tweets. To illustrate, an individual experiencing depression may consecutively share a series

to effectively capture both local contextual features and long-range dependencies [31,38].

Moreover, attention mechanisms [1,34,37] are employed to enable models to focus on the most salient aspects of the input. Additionally, multi-task learning is harnessed to jointly train models alongside other auxiliary tasks, such as statistical feature classification [35] and depression causation prediction [36], which yield supplementary insights for depression detection.

Recently, researchers have gathered data through online surveys and online social discourse, leveraging the substantial number of users and tweets on social platforms to obtain an ample and longitudinally sampled dataset. This approach effectively addresses the previously mentioned issue of collecting depression-related data. One study collected data from over 4,000 individuals, encompassing both depressed and non-depressed users on Twitter. [30] The dataset comprised over 2,000 samples from individuals diagnosed with depression, encompassing personal information and an anchor tweet to determine their depressive status. Additionally, all tweets posted within the 30 days preceding the anchor tweet were collected, based on the clinical characteristic of prolonged mood despondency among depression patients. For non-depressed users, relevant personal information and tweets within the same 30-day period were collected. However, this dataset did not account for the issue of data imbalance. Building upon this dataset, [39] further integrated personal information features and textual features, effectively removing redundant features through the utilization of k-means algorithm and the Bart summarization model [17], thereby improving the accuracy of depression identification. Our work builds upon the aforementioned research, focusing primarily on addressing data redundancy and data imbalance issues.

3 Method

In this section, we present our model for depression. Our model utilizes both the users' personal information and their historical posts as the foundation for detecting depression. Figure 1 illustrates our model diagram.

3.1 Task Definition

For social media datasets, users' posts often exhibit redundancy, irrelevance to their status, and may even contain unusable information, posing significant challenges for researchers to effectively extract user information. Herein, we have established the relevant symbols as defined in the article. We assume a user, denoted as U_i, has a total of n tweets in their history: $[T_1, T_2, \ldots, T_t, \ldots, T_n]$, where the t-th tweet represents the user's recent post. Our objective is to determine whether user U_i is a depression sufferer, for which we define the label as $y_i \in depression, undepression$. To achieve our goal, we amalgamate each user's profile information with their historical posts. We explore three approaches to process a user's historical tweets to address this issue: conducting direct tweet

extract activity information typically mandate real-time tracking for a duration exceeding two weeks. This extended timeframe, coupled with the requisite high level of patient cooperation, introduces significant costs and complexity into the research process.

Efforts to enhance the precision and applicability of AI-based depression identification must contend with these multifaceted challenges.

Our contributions are as follows:

1. We present a comprehensive depression assessment model that concurrently leverages users' historical tweets and their personal information. To achieve this, we meticulously devise distinct models for both historical tweets and user information. Our approach involves the integration of two discrete methodologies, one dedicated to handling multiple tweets and the other addressing the inherent challenge of imbalanced depression data.
2. We reprocess the depression tweet dataset to enhance its practical utility.
3. Through rigorous experimentation, we showcase the efficacy of our model in discriminating depression, yielding compelling empirical results and partially alleviating the associated challenges.

2 Related Work

In the field of psychology, early scholars have observed the theoretical correlations between mental health conditions and specific linguistic attributes, such as the presence of "depressive language" [2] advanced cognitive therapy and emphasized the significance of the frequency of negatively-valenced words, while other researchers [25] focused on the utilization of first-person pronouns and the patients' negative anticipations. Subsequent empirical investigations have validated these hypotheses and revealed associations between specific linguistic characteristics and the mental states of patients. Consequently, numerous studies utilize social media as a rich source of textual data, employing online user-generated posts for the manual analysis of mental health conditions. [7,26].

However, due to the burgeoning volume of online texts and the sensitivity of mental health conditions, manual text analysis and large-scale psychiatric interventions are no longer tenable. Consequently, Natural Language Processing (NLP) and text mining technologies have been harnessed to automate the analysis of mental health from social media data. While these approaches are not intended for definitive diagnoses, they do offer assistance in early detection [11,22,27]

Advancements in the realm of deep learning also bolster tasks related to mental health. The most recent methodologies employ deep learning models to automatically capture latent semantic information without the need for explicit feature engineering. Some studies utilize Convolutional Neural Networks (CNN) [33] or Recurrent Neural Networks (RNN) [8], including Long Short-Term Memory (LSTM) [12] and Gated Recurrent Unit (GRU) [28], to discern depression. Researchers also explore hybrid architectures combining CNN and RNN

1 Introduction

With the advancement and progress of society, people's material living standards are constantly improving, and psychological issues are receiving increasing attention. Psychological disorders are prevalent among young individuals, with approximately 75% of cases emerging during adolescence [15]. According to estimates by the World Health Organization, depression is one of the most prevalent psychological disorders, and by the year 2030, depression is projected to become the leading burden of disease globally [16,21].

Depression is characterized by significant and enduring mood melancholy, with symptoms encompassing sleep disturbances, appetite changes, and mental turmoil [3–5]. Despite its high prevalence, there is evidence indicating that 60% of severely depressed adolescents do not receive treatment.

Depression possesses a covert nature, and its occurrence is influenced by intricate factors such as heredity, gender, living environment, and physical ailments, rendering its diagnosis exceedingly challenging [4–6]. Presently, an accurate diagnosis of depression necessitates psychiatric practitioners to employ systematic inquiries, psychiatric examinations, and supplementary assessments, such as the Hamilton Rating Scale for Depression (HAMD) and the Patient Health Questionnaire-9 (PHQ-9) self-rating scale. Thus, the diagnostic evaluation heavily relies on patients' self-reported severity of depressive symptoms or clinical judgment regarding symptom severity. However, the advent of artificial intelligence-based approaches has presented the potential for objective diagnosis.

We focus on depression detection based on social networks. Recently, [14] discovered that a lack of social interaction increases the risk of depression. [24] analyzed the behavior and language usage of depressed users on Twitter. People's tweets on social networks such as Facebook, Twitter, and Weibo can be used to assess the risk of various mental health issues, such as depression and anxiety. [32] employed lemmatization tools to vectorize more recent tweets, reducing redundant features. [13] utilized a multimodal model and applied reinforcement learning to merge textual and image features of tweets, thereby improving the accuracy of depression identification [19].

The efficacy of depression identification through artificial intelligence remains suboptimal, encountering several noteworthy challenges. These challenges may be succinctly outlined as follows: **Limited Sample Size:** The recruitment of patients poses a substantial hurdle due to ethical concerns within the medical field. Consequently, a pervasive issue in depression studies is the constraint imposed by small sample sizes. This limitation complicates the attainment of definitive conclusions regarding individual-level depression diagnoses. **Data Complexity:** The datasets employed in these studies often exhibit a profusion of irrelevant features and noise. This characteristic not only augments the computational intricacy of algorithms but also compromises the predictive performance of models. Additionally, an inherent imbalance exists within the dataset, stemming from a lower representation of depression cases. **Temporal Constraints:** Extracting nuanced daily life characteristics of patients necessitates a protracted timeline. For instance, methodologies reliant on mobile devices to

Sentiment Analysis Based on Social Media - Early Stress and Depression Detection

Zixuan Li[1,3], Yuxuan Hu[1,3], Chenwei Zhang[1,3], Chengming Li[1,2(✉)], and Xiping Hu[1,2(✉)]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
{lizx76,huyx55,zhangshw7}@mail2.sysu.edu.cn, {licm,huxp}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen 518172, Guangdong, China
[3] Sun Yat-sen University, Shenzhen, Guangdong, China

Abstract. Depression has recently gained significant attention as a condition marked by persistent and profound mood disturbances. Extensive research suggests that depression can influence individuals' online speech behavior, manifested through the use of depressive language and a reduction in posting frequency. Our system seamlessly integrates various sources of information, including historical tweets and user profile data. Concerning historical tweets, we propose two methods to navigate the extensive and intricate user tweet history. Our findings indicate that these methods yield more pertinent user information. Subsequently, we input this information into our meticulously constructed deep learning classification model. This model is built upon a pre-trained BERT (Bidirectional Encoder Representations from Transformers) and a bidirectional LSTM (Long Short-Term Memory) model that incorporates attention mechanisms. In the context of user information, we extract relevant details and directly incorporate them into a deep learning model based on bidirectional GRU (Gated Recurrent Unit) and MLP (Multi-Layer Perceptron). Concurrently, to address the challenge of imbalanced depression datasets, we introduce Focal Loss and Dice Loss. Our experimental results underscore the effectiveness of these loss functions in our model. To validate the efficacy of our system, we reprocess the depression tweet dataset and conduct a series of experiments. Through these experiments, we conclusively demonstrate the robustness of our model, effectively mitigating the challenge of sample imbalance to a considerable extent.

Keywords: Deep learning · Social network · Depression recognition · Data imbalance

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 26–39, 2024.
https://doi.org/10.1007/978-3-031-65126-7_3

26. Perkins, T.K., Kern, L.R.: Widths of hydraulic fractures. J. Petrol. Technol. **13**(09), 937–949 (1961)
27. Nordgren, R.P.: Propagation of a vertical hydraulic fracture. Soc. Petrol. Eng. J. **12**(04), 306–314 (1972)
28. Hudson, J.A.: A critical examination of indirect tensile strength tests for brittle rocks [Ph. D. Thesis]. University of Minnesota, Minneapolis (1984)
29. Wang, C., Wang, R., Wang, C.: Development of multiple-diameter core hydraulic fracturing machine to test tensile strength of rocks. Chin. J. Rock Mech. Eng. **36**(S1), 3321–3331 (2017)
30. Cuisiat, F.D., Haimson, B.C.: Scale effects in rock mass stressmeasurements. Int. J. Rock Mech. Min. Sci. Geomech. Abst. **29**(2), 99–117 (1992)
31. Hou, B., Chen, M., Wan, C., Sun, T.: Laboratory studies of fracture geometry in multistage hydraulic fracturing under triaxial stresses. Chem. Technol. Fuels Oils **53**(2), 219–226 (2017)
32. Park, J.Y., Tuell, G.: Conceptual design of the CZMIL data processing system (DPS): algorithms and software for fusing lidar, hyperspectral data, and digital images. Proc Spie **7695**(5), 731–739 (2010)
33. Qin, H.: Constructions of uniform designs with mixed levels. Acta Math. Appl. Sin. **28**(4), 704–712 (2005)
34. Montgomery, D.C., Peck, E.A.: Introduction to linear regression analysis (1982)
35. Schmitt, D.R., Zoback, M.D.: Diminished pore pressure in low-porosity crystalline rock under tensional failure; apparent strengthening by dilatancy. J. Geophys. Res. **97**(B1), 273–288 (1992)
36. Ito, T., Satoh, T., Kato, H.: Deep Rock Stress Measurement by Hydraulic Fracturing Method Taking Account of System Compliance Effect. Xie Furen. CRC Press, Boca Raton (2010)
37. Zhu, X., Zhang, J., Feng, J.: Multiobjective particle swarm optimization based on PAM and uniform design. Math. Probl. Eng. **2015**, 1–17 (2015)
38. Yang, L., Pan, F., Weifeng, J., Shengwei, S., Yong, Z., Tao, Z.: Predictive method of nonlinear system based on artificial neural network and svm. Oxidat. Commun. **39**(1Appa), 1226–1235 (2016)

4. Haimson, B.C.: Hydraulic Fracturing in Porous and Nonporous Rock and its Potential for Determining In Situ Stresses at Great Depth. University of Minnesota, Minneapolis (1968)
5. Haimson, B.C., Fairhurst, C.: Initiation and extension of hydraulic fractures in rocks. Soc. Petrol. Eng **9**, 310–318 (1967)
6. Von Schonfeldt, H., Fairhurst, C.: Field experiments on hydraulic fracturing. Soc. Petrol. Eng. AIME **12**(1), 69–77 (1972)
7. Wang, C.: Brief review and outlook of main estimate and measurement methods for in-situ stresses in rock mass. Geol. Bull. China **60**(5), 971–996 (2014)
8. Wang, J., Li, H., Wang, Y., Li, Y., Jiang, B., Luo, W.: A new model to predict productivity of multiple-fractured horizontal well in naturally fractured reservoirs. Math. Probl. Eng. **2015**, 1–9 (2015)
9. Xie, F., Chen, Q.: Study on the Crustal Stress Environment in China. Geological Press, Beijing (2003)
10. Jaeger, J.C., Cook, N.G.W., Zimmerman, R.W.: Fundamentals of Rock Mechanics. Blackwell Publishing, London (2007)
11. Wang, C., Song, C., Xing, B.: Compliance of drilling-rod system for hydro-fracturing in situ stress measurement and its effect on measurements at great depth. Geoscience **26**(4), 808–816 (2012)
12. Zoback, M.D., Pollard, D.D.: Hydraulic fracture propagation and the interpretation of pressure-time records for in-situ stress determinations. In: 19th US Symposium on Rock Mechanics (USRMS), pp. 14–22. American Rock Mechanics Association (1978)
13. Ito, T., Hayashi, K.: Physical background to the breakdown pressure in hydraulic fracturing tectonic stress measurements. Int. J. Rock Mech. Mining Sci. Geomech. Abst. **28**(4), 285–293 (1991)
14. Chang, C., Jo, Y., Oh, Y., Lee, T.J., Kim, K.: Hydraulic fracturing in situ stress estimations in a potential geothermal site, Seokmo Island, South Korea. Rock Mech. Rock Eng. **47**(5), 1793–1808 (2014)
15. Zhou, L., Ding, L., Guo, Q.: Experimental study of absolute rock stress measurements under different fracture media. Rock Soil Mech. **10**, 2869–2876 (2013)
16. Zhang, J.: Analysis of the Hydromechanics Factors Impact on Hydraulic Fracturing In-situ Stress Measurement. China University of Geosciences, Beijing (2018)
17. Matsunaga, I., Kobayashi, H., Sasaki, S.: Studying hydraulic fracturing mechanism by laboratory experiments with acoustic emission monitoring. Int. J. Rock Mech. Min. Sci. Geomech. Abst. **7**, 909–912 (1993)
18. Ishida, T., Chen, Q., Mizuta, Y.: Effect of injected water on hydraulic fracturing deduced from acoustic emission monitoring. In: Seismicity Associated with Mines, Reservoirs and Fluid Injections. Birkhäuser, Basel (1997)
19. Fang, K.: Uniform Design and Uniform Design Table. Science Press, Beijing (1994)
20. Wang, Z., Fang, K.: Measures of uniformity for uniform designs with qualitative factors. Math. Stat. Manag. **19**(3), 28–32 (2000)
21. Myers, R.H.: Classical and Modern Regression with Applications, 2nd edn. Duxbury Press, Belomont (1994)
22. Liu, Y., Wang, C., Wang, J., Ji, W.: Optimization research on thermal error compensation of FOG in deep mining using uniform mixed-data design method. Math. Probl. Eng. **2019**, 1–6 (2019)
23. Zhang, L., Cai, X.: Uniformity masks design method based on the shadow matrix for coating materials with different condensation characteristics. Sci. World J. **2013**, 1–4 (2013)
24. Khristianovich, S.A., Zheltov, Y.P.: Formation of vertical fractures by means of a highly viscous fluid. In: Proceedings 4th World Petroleum Congress, pp 579–586 (1955)
25. Geertsma, J., De Klerk, F.: A rapid method of predicting width and extent of hydraulically induced fractures. J. Petrol. Technol. **21**(12), 1571–1581 (1969)

experimental results. Meanwhile, neural networks and deep learning algorithms are also considered to analyze and predict rock fracturing values, verifying the accuracy of the rock fracturing model [38].

6 Conclusions

In this paper, we utilized the optimal uniform design method to optimize hydraulic fracturing simulation experiments. The results showed that this design method not only reduced the number of experiments but also improved the uniformity and test effect. It provides a fast and effective way to develop an error compensation formula for various influence elements, aiming to enhance measurement accuracy of the hydrofracture method of geostress measurement.

- The paper proposed an optimal approach for the hydrofracture method of geostress measurement using the uniform design method. Considering these unique properties of the drilling mud, which involves multiple hydraulic elements and values. This paper constructs an experimental plan that can simplify the testing procedure, and decrease implementation fees. As a result, the efficiency of the simulation hydraulic fracturing tests can be significantly enhanced.
- This study examined the impact of various hydrodynamic factors on rock fracturing pressure using test results. Through multivariate regression analysis, an optimal regression model for multiple influencing factors (fracturing fluid viscosity, density, axial load, and injection rate) was obtained. Additionally, the values of instantaneous rock splitting were discussed in depth, and the validity of Perkins-Kern-Nordgren (PKN) classical mechanical model in theoretical analysis was confirmed.

Acknowledgments. This work was supported by Project funded by National Natural Science Youth Foundation of China (41804089), and Geological survey Project of China Geological Survey (DD20230447).

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 there are no conflicts of interest regarding the publication of this paper, and the authors confirm that the mentioned received funding in the "Acknowledgment" section did not lead to any conflict of interests regarding the publication of this manuscript.

References

1. Clark, J.B.: A hydraulic process for increasing the productivity of wells. J. Petrol. Technol. **1**(1), 1–8 (1949)
2. Zhao, Z., Guo, J., Ma, S.: The Experimental investigation of hydraulic fracture propagation characteristics in glutenite formation. Adv. Mater. Sci. Eng. **2015**, 1–5 (2015)
3. Hubbert, K.M., Willis, D.G.: Mechanics of hydraulic fracturing. Trans. AIME **210**(1), 153–168 (1957)

To summarize, Eqs. (6) and (7) demonstrate a similar changing pattern as x_1 and x_4, but undergoes an opposite changing pattern in relation to x_2 (fracturing fluid density). The simulation experiments confirmed a strong correlation between rock fracturing pressure and the viscosity, density, and injection rate of the fracturing fluid. These findings support the conclusions of theoretical analysis in Sect. 2. This indicates a good distribution uniformity of experimental points. Additionally, this study validated the efficiency and suitability of the experimental method in establishing a fracture pressure correction formula for various hydrodynamic factors.

5 Discussion

Simulation experiments were conducted to analyze the impact of various factors such as injection rate, density, viscosity, and fracturing fluid medium on hydraulic fracturing. These experiments provide a fast and reliable way to understand the influence of hydrodynamic factors on hydraulic fracturing. A compensation model can be utilized to minimize the interference of hydrodynamic factors and enhance the accuracy of in-situ stress measurement during hydraulic fracturing in practical engineering applications. Consequently, simulation tests have the potential to improve the measurement accuracy of hydraulic fracturing methods. However, this study only considered three fluid mechanics parameters, namely injection rate, fracturing fluid density, and viscosity. Therefore, future research should explore the incorporation of additional hydrodynamic parameters such as hydraulic friction, liquid compressibility, and the effects of different types of fracturing fluid media on the effective fracturing pressure of rocks. These insights are valuable in advancing our understanding of hydraulic fracturing in practical applications.

- In terms of different fracturing fluids, such as hydraulic oil, mud, and aqueous solution, test results from Zhou Longshou (2013) [15] and Zhang Jie, Wang Chenghu et al. (2017) [16] indicate that mud and hydraulic oil lead to higher rock fracturing pressures compared to aqueous solutions. The combination of densities and viscosities of the fracturing fluids greatly affects the rock fracture pressure, while the compressibility of the fracturing fluid also influences the flexibility of the hydraulic fracturing measurement system (Wang Chenghu et al., 2012) [11], thereby affecting the measurement results. This study only considered mud as the fracturing fluid, so future studies should include more representative hydrodynamic parameters and different types of fracturing fluids for a comprehensive analysis of their influence on rock fracture pressure. Additionally, an appropriate correction formula and compensation model should be established for hydraulic fracturing errors under different working conditions.
- The experimental results confirmed that the injection rate of the fracturing fluid has a significant impact on the rock fracturing pressure, with a proportional increase. This finding aligns with the results of hydraulic fracturing tests conducted by several foreign researchers (Ito and Hayashi, 1991; Schmitt et al., 1992; Zo-back et al., 2007) [12–14, 36, 37]. To further enhance the accuracy of future simulation tests and reduce losses associated with hydraulic friction, especially head loss, it is recommended to install a high-precision pressure sensor in the fracturing test section. This enhancement will allow for a better analysis of the influence of injection rate on the

Table 4. Multi-factor VIF value

VIF	Value
x_1	1.0027
x_2	1.0009
x_3	1.0035
x_4	1.0021

polynomial. The Matlab Linear-Model.stepwise function was utilized to perform a multivariate quadratic polynomial regression of the data presented in Table 4. Based on this regression analysis, the stepwise regression method of the LinearModel class object was employed to establish the Eq. (6), to show the relationship between factors (x_1, x_2, x_3, x_4) and fracturing value (y).

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_3^2 + b_6x_1x_4 + b_7x_4^2 \tag{6}$$

Based on the stepwise regression calculation results, the resulting equation for multivariate polynomial regression can be expressed as follows:

$$y = 17.937 + 0.023x_1 - 10.266x_2 + 2.054x_3 + 1.598x_4 - 0.361x_3^2 - 0.067x_1x_4 + 18.535x_4^2 \tag{7}$$

Furthermore, Eq. (8) produces a $p\text{-}value_2 = 0.000405$, and $p\text{-}value_2 << 0.05$ (significance level). Figure 3 illustrates the regression fitting plots of Eqs. (6) and (7), demonstrating a higher degree of fit in the latter. Therefore, Eq. (7) is an optimal fitting formula for (y) and (x_1, x_2, x_3, x_4) in this design.

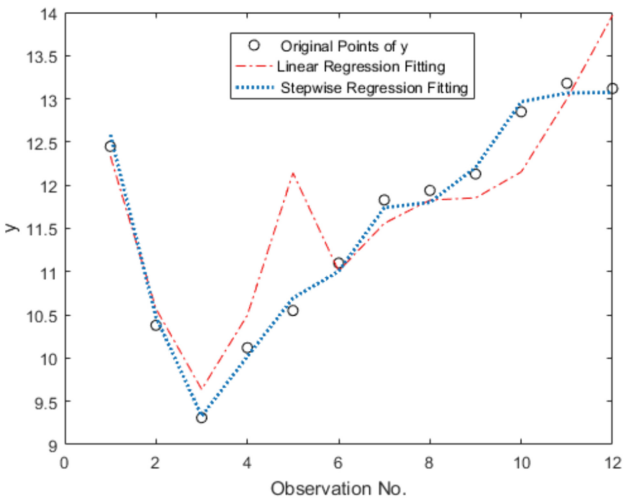


Fig. 3. The optimal fitting formula for linear regression and stepwise regression fitting

4 Results Analysis

4.1 Experimental Results

Table 3 presents results of the experiments using the uniform design method involved in Sect. 3, showcasing the obtained effective rock fracturing value.

Table 3. Table of hydraulic fracturing values of the simulation experiments

No.	Influencing factors				Results
	Viscosity(g/cm^3)	Density(mPa•s)	Axial Compression(MPa)	Injection Rate(MPa/s)	Fracturing pressure(MPa)
1	280	1.4	1.2	0.1	12.46
2	170	1.6	3.6	0.1	10.35
3	70	1.6	1.2	0.55	9.34
4	70	1.2	2.4	0.05	10.12
5	70	1.0	4.8	0.1	10.58
6	150	1.2	4.8	0.4	11.12
7	280	1.6	4.8	0.1	11.85
8	130	1.2	2.4	0.2	11.91
9	150	1.2	3.6	0.2	12.1
10	170	1.2	2.4	0.4	12.88
11	170	1.0	1.2	0.05	13.19
12	280	1.0	3.6	0.55	13.15

4.2 Multivariate Polynomial Regression

Regression analysis is a method used to establish the relationship between the dependent variable y and the independent variables (x_1, x_2,…, x_i) [34–36]. In Eq. (6), y represents actual demonstration value of the rupture pressure, x_1 represents the density, x_2 represents the viscosity, x_3 represents the axial pressure, and x_4 represents the injection speed. Table 4 underwent multiple linear regression and multivariate polynomial regression to determine the respective fitting models. These models were then compared to obtain the optimal fitting formula.

A regression model was subjected to a multicollinearity diagnosis using the variance inflation factor (VIF) method, which resulted in Table 4. Generally, if VIF < 5, there is no collinearity. The independent variables in Table 4 had VIFs below 5, indicating the absence of multicollinearity in the model.

In order to enhance the non-linear terms in the model, a stepwise regression approach was employed to conduct a generalized linear regression analysis using a quadratic

Table 1. Elements and their numerical value

Element	Level	Parameter value
Viscosity	4	70; 150; 170; 280
Density	4	1.0; 1.2; 1.4; 1.6
Axial compression	4	1.2; 2.4; 3.6; 4.8
Injection rate	6	0.05; 0.1; 0.2; 0.25; 0.4; 0.55

In order to account for the numerous elements and their values, using mud as fracturing fluid, an optimized experimental scheme based on mixed-level uniform was developed. Using the Data Processing System (DPS) software [33], a total of 12 experiments were conducted, as shown in Table 2. The constructed optimal mixed-level uniform design table $U_{12}^{*}\left(6 \times 4^{3}\right)$ was subjected to a maximum of 1000 iterations.

Table 2. Table of Influencing elements in $U_{12}^{*}\left(6 \times 4^{3}\right)$

No.	Influencing elements			
	x_1	x_2	x_3	x_4
1	4	3	1	3
2	3	4	3	3
3	1	4	1	5
4	1	3	2	1
5	1	1	4	3
6	2	3	4	6
7	4	4	4	2
8	2	2	2	4
9	2	2	3	1
10	3	2	2	6
11	3	1	1	2
12	4	1	3	5

A smaller D implies better uniformity of the experimental design [19, 20]. By calculating the Eq. (6), we obtained that $D^{*} = 0.1713$ for the $U_{12}^{*}\left(6 \times 4^{3}\right)$, which exhibits a good distribution uniformity.

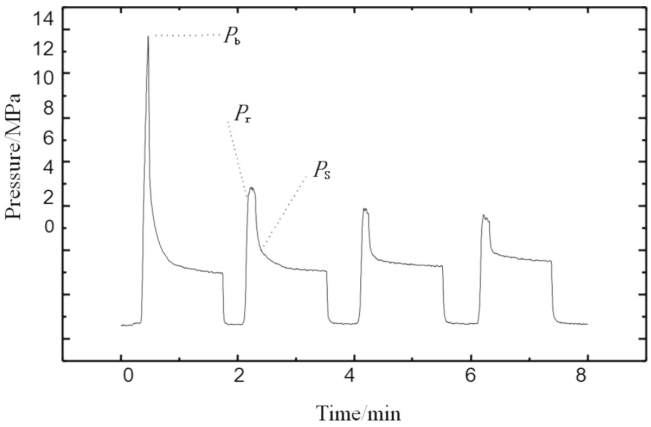


Fig. 2. Typical Pressure-Time Record Curve in Hydraulic Fracturing

and the pressure value at this time is recorded as the instantaneous closure pressure P_s. After releasing the pressure, reloading causes the fracture to reopen, and the pressure value at this time is recorded as the reopening pressure P_r.

According to the elastic theory and the PKN mechanical model, as shown in Fig. 3, the fracturing pressure of the rock in the fracturing section is:

$$P_b = 3\sigma_h - \sigma_H + T \tag{5}$$

Among them, σ_H and σ_h are the maximum and minimum horizontal principal stresses, respectively, and T is the tensile strength of the rock. The fractures induced by hydraulic fracturing are vertical fractures and perpendicular to the direction of the minimum horizontal principal stress. Equation (5) indicates that the fracturing pressure of rocks is independent of the size of the borehole and the elastic modulus of the rock, and is mainly determined by the tensile strength of the rock and the magnitude of the in-situ stress around the borehole.

3 Optimal Design of the Testing

The high pressure fluids are commonly applied in hydraulic fracturing simulation experiments, including clean water, hydraulic fluid, carboxymethyl cellulose (CMC) aqueous solution, and drilling mud [30–32]. The density and viscosity of the mud medium can be adjusted according to the requirements of the simulation experiment. For these fracturing fluid media, only a small number of factors and tests are required, so conventional comprehensive experimental methods can be used for their respective simulation experiments. In contrast, there are more parameters and theirs values in the mud medium, so an optimal design based on uniform design method is suitable for the testing.

The theoretical analysis of the PKN model revealed that when using mud as the fracturing fluid medium in the simulation test, it requires three hydrodynamic factors: density, injection rate, and viscosity, as well as a factor of loading axial compression. Different numerical values of each element are presented in Table 1.

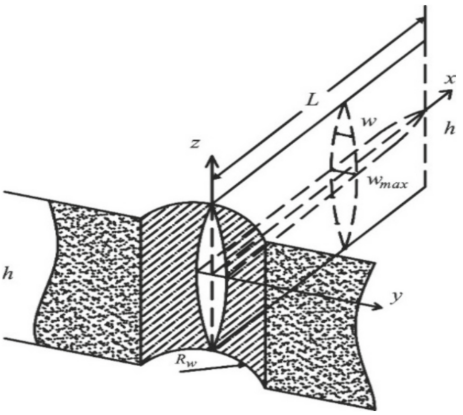


Fig. 1. PKN classical mechanical model

Here, *q(x, t)* represents the volume of fluid flowing through the cross-section of the crack, $q_l(x, t)$ represents the volume of fluid lost per unit length of the crack, and *A(x, t)* represents the cross-sectional area of the crack. When there is no fluid leakage, the length of the crack L, its local width w, and the pore pressure P_w can be calculated [28, 29]:

$$L = 0.68\left[\frac{GQ^3}{(1-\upsilon)\mu h^4}\right]^{\frac{1}{5}} t^{\frac{4}{5}} \tag{2}$$

$$w = 2.5\left[\frac{(1-\upsilon)\mu Q^2}{Gh}\right]^{\frac{1}{5}} t^{\frac{1}{5}} \tag{3}$$

$$p_w = 2.5\left[\frac{G^4\mu Q^2}{(1-\upsilon)^4 h^6}\right]^{\frac{1}{5}} t^{\frac{1}{5}} \tag{4}$$

The following variables are used in this context: G (shear modulus), υ (Poisson ratio), h (length), Q (injection rate), and μ (viscosity).

2.2 Principles of Hydraulic Fracturing Measurement

The basic principle of in-situ stress measurement based on hydraulic fracturing involves placing drill rods and packers into a borehole using a drilling rig to measure their positions. Fluid is injected into the packers through a loading control system, isolating a test section within the borehole, and the fluid is further injected into the test section until fracturing occurs.

As shown in Fig. 2, the first highest pressure value is recorded as the fracturing pressure P_b. Then the pressure drops rapidly to a state of fluid seepage into the fracture and remains constant. At this point, the pump is turned off to stop loading, and the pressure in the fracturing section decreases rapidly, causing the fracture to close quickly. When the fracture is in the near-closed state, the rate of pressure decrease slows down,

of error include: (1) the drill pipe and the packer deformation [9–11]; (2) variations in the determination method of the measurement curve during data analysis [12]; and (3) different category and the associated performance factors of the fracturing fluids [13–17].

Many researchers have dedicated themslves to explore the fracturing fluids impact on rock fracturing. For instance, Ito (1991) and Chang (2014) suggested that increasing the injection rate of fracturing fluid and considering factors like flow rate, viscosity, and density can enhance the tensile strength of the rock. Zhou et al. (2013) and Zhang (2018) conducted tests using different density mud media as fracturing fluids and observed significant variations in rock fracturing behavior. Matsagaga (1993) and Ishida et al. (1997) verified the impact of fracturing fluid viscosity on rock fracturing through oil drilling experiments. Wang (2012) and Zhou (2013) analyzed the error in stress measurement caused by the compressibility of clear water used as a fracturing fluid and its effect on system flexibility. These studies contribute to a better understanding of fluid mechanics factors in accurate rock fracturing measurements.

In summary, the hydrodynamic factors that influence rock fracturing during hydraulic fracturing include flow velocity, viscosity, density, and compressibility. Conducting simulation experiments based on these factors is crucial for understanding their impact on rock fracturing. However, these experiments can be destructive to the testing core, making them complicated and costly to design. To address this, the uniform design method has been proposed as an experimental design approach that evenly spreads test points throughout the range of variables, requiring fewer trials compared to other methods [19, 20]. In particular, the design aims to conduct trials with many experimental factors and a large number of levels, with fewer trials required compared with the orthogonal design or comprehensive design methods [21–23]. In this study, a uniform table for experiment design is used to combine selected hydrodynamic factors of the fracturing fluid with the factor of horizontal pressure. This approach reduces the test times while ensuring their effectiveness and significantly improving efficiency. The results of these experiments are then analyzed to determine the effects of the factors on rock fracturing value.

2 Error Analysis of Hydraulic Fracturing Theory

2.1 PKN Mechanical Model

The borehole used to measure hydraulic fracturing in-situ stress was typically vertical and primarily influenced by the maximum horizontal principal stress, and the minimum stress, and minimum is same as it is [24]. The fracturing crack was vertical because it was perpendicular to the minimum horizontal principal stress plane [25, 26]. The authors used the PKN classical mechanical model [27, 28] to analyze how fluid mechanics affects the fracture crack and its fracture pressure in this paper.

Figure 1 illustrates the PKN fracturing crack model [27, 28]. Nordgren (1972) obtained the fluid's continuity equation in the crack, ignoring the compression properties of the fracturing fluid [28]:

$$\frac{\partial q}{\partial x} + q_t + \frac{\partial q}{\partial t} = 0 \tag{1}$$

Optimal Design of Hydraulic Fracturing Simulation Experiments for In-Situ Stress Measurement

Yang Li[1], Daji Zhang[3], and Yimin Liu[2(✉)]

[1] Institute of Exploration Technology, CGS, Chengdu 611734, China
[2] Chengdu Technological University, Chengdu 611730, China
153973418@qq.com
[3] Chengdu Rail Transit Group Co., Ltd., Chengdu 610036, China

Abstract. In order to account for a large number of hydrodynamic influencing factors with multiple levels in rock fracturing experiments, the uniform design method is frequently utilized instead of conventional methods like comprehensive and orthogonal designs, as they significantly impact the experimental effects. Based on the Perkins-Kern-Nordgren (PKN) model, the influencing factors of injection rate, viscosity, and density of the fracturing fluid, along with their corresponding parameter values or levels, were taken into consideration to construct an optimal table $U_{12}^{*}\left(6 \times 4^{3}\right)$ for experiment design. Subsequently, an optimized experimental scheme was developed. The experimental results based on this design were analyzed using multiple regression analysis to establish an optimal regression equation for the influencing factors (x_1, x_2, x_3, x_4, representing fluid viscosity, density, loading axial compression, and injection rate, respectively) and to determine the corresponding rock fracturing value (y). This indicates a good distribution uniformity of experimental points. Additionally, this study validated the efficiency and suitability of the experimental method in establishing a fracture pressure correction formula for various hydrodynamic factors, and it is also a precise approach for geostress measurements.

Keywords: Uniform design method · Mixed-level · In-situ stress measurement · Rock fracturing

1 Introduction

The stress stored in the interior of a rock mass without disturbance is referred to as in-situ stress, which has multiple sources and is influenced by various factors, resulting in a complex and variable distribution of stress in the Earth's crust [1]. Hydraulic fracturing is a crucial technique for measuring in-situ stress in various geological structures such as hydropower stations, tunnels, chambers et al. [2–4]. This approach offers an efficient test procedure, along with straightforward data analysis procedures. However, several factors can influence the accuracy of rock fracturing measurements [5–7]. The primary sources

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 15–25, 2024.
https://doi.org/10.1007/978-3-031-65126-7_2

33. Xia, R., Liu, Y.: A multi-task learning framework for emotion recognition using 2d continuous space. IEEE Trans. Affect. Comput. **8**(1), 3–14 (2017). https://doi.org/10.1109/TAFFC.2015.2512598
34. Xie, B.: Research on key technology of Mandarin speech emotion recognition. Ph.D. thesis, Zhejiang University (2006)
35. Xu, Y., Su, H., Ma, G., Liu, X.: A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context. Complex Intell. Syst. **9**(1), 951–963 (2023)
36. Yu, W., et al.: Ch-sims: a chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3718–3727 (2020)

16. Li, Y., Zhao, T., Kawahara, T., et al.: Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In: Interspeech, pp. 2803–2807 (2019)
17. Liu, A.T., Yang, S.W., Chi, P.H., Hsu, P.C., Lee, H.V.: Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6419–6423. IEEE (2020)
18. Liu, Y., et al.: Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and AV-Mixup consistent module. In: Proceedings of the 2022 International Conference on Multimodal Interaction, pp. 247–258 (2022)
19. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north American English. PLoS ONE **13**(5), e0196391 (2018)
20. Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In: Ismir, vol. 270, p. 11. Plymouth, MA (2000)
21. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F.: Multimodal emotion recognition on ravdess dataset using transfer learning. Sensors **21**(22), 7665 (2021)
22. McFee, B., et al.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25 (2015)
23. Nwe, T.L., Foo, S.W., De Silva, L.C.: Classification of stress in speech using linear and nonlinear features. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP 2003), vol. 2, pp. II–9. IEEE (2003)
24. Oh, K.J., Lee, D., Ko, B., Choi, H.J.: A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In: 2017 18th IEEE International Conference on Mobile Data Management (MDM), pp. 371–375. IEEE (2017)
25. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
26. Pravena, D., Govind, D.: Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. Int. J. Speech Technol. **20**(4), 787–797 (2017)
27. Singh, A., Liu, H., Plumbley, M.D.: E-panns: sound recognition using efficient pre-trained audio neural networks. arXiv preprint arXiv:2305.18665 (2023)
28. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21–25. IEEE (2021)
29. Triantafyllopoulos, A., Schuller, B.W.: The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7268–7272. IEEE (2021)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. **30** (2017)
31. Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., Ma, C., Huang, Y.: Video sentiment analysis with bimodal information-augmented multi-head attention. Knowl.-Based Syst. **235**, 107676 (2022)
32. Xi, Y., Li, P., Song, Y., Jiang, Y., Dai, L.: Speaker to emotion: domain adaptation for speech emotion recognition with residual adapters. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 513–518. IEEE (2019)

5 Conclusions

In this paper, we effectively applied transfer learning to fine-tune the English pre-trained model, achieving a notable improvement in the F1 score to 0.46, significantly surpassing the baseline of 24%. For future research, we will continue exploring the cross-corpus SER domain and further investigating other deep learning techniques to enhance the performance of the transfer learning models in emotion recognition.

References

1. Amiriparian, S., et al.: Snore sound classification using image-based deep spectrum features (2017)
2. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335–359 (2008)
3. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
4. Gemmeke, J.F., et al.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 776–780. IEEE (2017)
5. Gong, Y., Chung, Y.A., Glass, J.: Ast: audio spectrogram transformer. arXiv preprint arXiv:2104.01778 (2021)
6. Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: Ssast: self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 10699–10709 (2022)
7. Haq, S., Jackson, P.J.: Multimodal emotion recognition. In: Machine Audition: Principles, Algorithms and Systems, pp. 398–423. IGI Global (2011)
8. Hossain, M.S., Muhammad, G., Song, B., Hassan, M.M., Alelaiwi, A., Alamri, A.: Audio-visual emotion-aware cloud gaming framework. IEEE Trans. Circuits Syst. Video Technol. **25**(12), 2105–2118 (2015)
9. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 801–804 (2014)
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894 (2020)
12. Koolagudi, S.G., Murthy, Y.S., Bhaskar, S.P.: Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. Int. J. Speech Technol. **21**(1), 167–183 (2018)
13. Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., Stober, S.: Transfer learning for speech recognition on a budget. arXiv preprint arXiv:1706.00290 (2017)
14. Lee, M.C., Chiang, S.Y., Yeh, S.C., Wen, T.F.: Study on emotion recognition and companion chatbot using deep neural network. Multimedia Tools Appl. **79**, 19629–19657 (2020)
15. Li, P., Song, Y., McLoughlin, I.V., Guo, W., Dai, L.R.: An attention pooling based representation learning method for speech emotion recognition (2018)

Table 1. Quantitative evaluation of the different strategies on speech emotion recognition. In bold, the best model.

Classification	train F1 score	val F1 score	train Acc	Val Acc
original frozen CNN10(baseline)	0.2760	0.3125	0.3658	0.3994
original frozen CNN14(baseline)	0.2528	0.2413	0.3644	0.3778
original CNN10	0.6612	0.4432	0.6831	0.4657
original CNN14	0.8822	0.4542	0.8901	**0.4774**
multihead CNN10	0.6297	0.4536	0.6576	0.4684
multihead CNN14	0.9144	0.4642	0.9208	0.4674
multilayer CNN10	**0.7694**	**0.4636**	**0.7871**	0.447
multilayer CNN14	**0.9516**	**0.4652**	**0.9613**	**0.4722**

also observed a large performance gain for valence and a lesser gain for other aspects. The results suggest that while fine-tuning does incur additional computational costs, the benefits it yields in terms of improved performance make it a worthwhile endeavor. The validation set F1 scores for both CNN10 and CNN14 models, when employing the multilayer multi-head attention module, surpass those of the baseline, with the CNN14 model also demonstrating higher accuracy on the validation set. A comparison between different structures reveals that the multilayer multi-head attention modules generally outperform their single-layer counterparts. Specifically, the 'multilayer CNN14' model delivered the best results, achieving optimal performance with the least amount of epochs.

4.4 Future Work

Compared to the baseline, we believe there is ample room to improve the accuracy of the validation set. There are several areas for future improvements. First, we did not adjust the architecture of the pre-trained model, and the limited number of CNN layers may have hindered its ability to recognize emotions arising from emotional correlations in the data fully. Thus, further adjustments to the model architecture and hyperparameters are necessary for better generalization. In addition, we should further explore the linguistic and cultural differences in the datasets. Our target dataset is in Mandarin Chinese, while the baseline dataset is in English. Cross-language disparities may impede significant performance improvements.

Revealing these potential differences between languages and cultures requires further research in multi-task learning and exploring the fields of language and cultural studies. These areas offer significant potential for future research efforts, helping bridge the cross-linguistic gap and improving the performance of deep learning algorithms in specific tasks.

attention layer is passed into the next layer, it first goes through an additional transformation via a fully connected layer. The potential benefit of this could be to provide an additional means to capture and transform more complex patterns in the data.

4 Experiment

4.1 Dataset Setup

In our experiments, we utilized the CH-SIMS v2.0 [18] dataset, which is partitioned into three sets: the training set (80%), the test set (10%), and validation set (10%). The obtained output is categorized into five distinct labels. For feature extraction, the librosa library [22] is employed to extract log mel spectrograms from raw audio data.

4.2 Experimental Setting

In the training experiments, we leverage a pre-trained model on the AudioSet dataset to facilitate transfer learning on an existing dataset. During the fine-tuning phase, we employed both single and triple multi-head self-attentive layers, with the results being labeled as 'multihead' and 'multilayer' respectively. The training process was utilized the Adam optimizer [10] and cross-entropy loss with a batch size of 16.

Results from the two original models (CNN10 and CNN14), with frozen parameters, were served as the baseline for our benchmark. In the fine-tuning phase, the models with multi-head and multi-layer were trained for 200 epochs with an initial learning rate of 1e-4. Each experiment set were conducted ten times with the average results recorded. The best-performing model is selected and saved, conducting experiments on both test sets and validation sets. The recorded results are presented in Table 1.

4.3 Results and Discussion

Table 1 presents the results of experiments conducted on the CH-SIMS2.0 dataset, with the primary evaluation metrics being the F1 score and accuracy (Acc). Remarkably, the fine-tuned models consistently outperform their counterparts with frozen parameters. When compared to other models, our approach delivers highly competitive results. The findings indicate that fine-tuning of parameters significantly enhances the accuracy of audio classification. Therefore, we firmly advocate for implementing parameter fine-tuning as an effective strategy to elevate output performance.

Table 1 provides a summary of the performance exhibited by the various speech emotion recognition models that were tested. When considering the experiments with the frozen initial parameters as the baseline, improvements are observed across all tested results in comparison to the baseline. Notably, we

every convolutional layer, and then ReLU non-linearity is applied to facilitate better training convergence. For CNN10 and CNN14, the convolutional blocks are used in pairs before an average pooling layer is applied. Specifically, CNN10 is composed of 8 convolutional blocks (4 pairs), while CNN14 consists of 12 convolutional blocks (6 pairs). All networks include a penultimate fully connected layer to enhance representation capability, succeeded by a final fully connected layer with 527 units. A sigmoid activation function is applied at this stage to derive the probabilities of each class.

3.3 Multi-head Attention Block

To capture the semantic relevance embedded within the speech signal, the multi-head self-attention mechanism [30] is employed to focus on emotional information from various subspaces. In the multi-head attention mechanism, there are H parallel attention heads, and each of these attention heads calculates a set of attention weights:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q^\top \cdot K}{\sqrt{d_k}}\right) \cdot V^\top \tag{1}$$

where: Q, K, and V are the query, key, and value matrices for calculating the multi-attention mechanism. The Softmax function is commonly used to normalize the attention scores and ensure that they represent a valid probability distribution, where the sum of all attention weights is equal to 1.

We use the optimization algorithm Noam for learning rate tuning to achieve better model solutions. By computing similarities between Q and K, the mechanism assigns weights to each query position, determining the significance of the corresponding values.

$$lr = factor \cdot modelsize^{-0.5} \cdot \min(step^{-0.5}, step \cdot warmup^{-1.5}) \tag{2}$$

where: *factor* refers to the initial learning rate size, *model size* denotes the hidden layer dimension, *step* represents the number of optimization steps, and *warmup* denotes the value of the step when the learning rate reaches its maximum value.

Between each layer, we introduced a gate function incorporating the sigmoid function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

x is the input variable. It converts the output of the model into a probability value between 0 and 1. This gated mechanism helps the model to dynamically adjust the importance of different layers and enables more flexible and adaptive information processing. Furthermore, the use of the sigmoid function ensures a smooth gating operation, avoiding abrupt changes and maintaining stability during the learning process.

After applying several layers of multi-head attention, we performed an fully connected layer (comprised of a linear layer and a ReLU activation function) following each gating mechanism. This means that before the output from each

3.1 CH-SIMS Dataset

To conduct Speech Emotion Recognition (SER) on Chinese speech, we utilized the CH-SIMS v2 training dataset [18]. This dataset comprises 60 original videos, resulting in 2,121 video segments. It offers a diverse range of character backgrounds, covering different age groups, and featuring high-quality recordings. Only Mandarin Chinese speech is included in this dataset.

Compared to version 1.0 [36], the video data in this updated version includes a broader range of scenarios, and the focus is on acoustic and visual features rather than text, encompassing a wider variety of emotional expressions. This aspect serves as a valuable inspiration for our research.

Each video segment in the CH-SIMS v2 dataset has undergone multimodal annotations, further categorized into five emotion categories:

$$\begin{aligned} \text{negative} &: \{-1.0, -0.8\}, \\ \text{weakly negative} &: \{-0.6, -0.4, -0.2\}, \\ \text{neutral} &: \{0.0\}, \\ \text{weakly positive} &: \{0.2, 0.4, 0.6\}, \\ \text{positive} &: \{0.8, 1.0\}. \end{aligned}$$

3.2 Pre-trained Block

Our approach aims to leverage a pre-trained speech recognition network to extract meaningful features from the samples of CH-SIMS. The CNN architectures utilized in our study are adapted from those presented in reference [11]. The PANNs framework houses a diverse of pre-trained models, encompassing various versions of CNN models. These models are trained on extensive audio datasets, empowering them with ability to capture intricate audio feature representations. This capability allows PANNs to efficiently capture and analyze patterns and recognizable features within audio data. We applied its subsample since, within PANNs [11], the CNN-14 model achieves the best performance, and also uses the pre-trained model corresponding here. Following the preprocessing phase, the vocal data is fed into the framework which then internally constructs a frequency-based representation of the recordings. Interestingly, in a related study [27], it was observed that CNN-10 model performs well with some smaller datasets. Consequently, in our experiments, we employed both CNN-10 and CNN-14 models for the feature extraction and embedding.

The audio data undergoes the following preprocessing steps: first, the audio is resampled to 32kHz. Then, a Short-Time Fourier Transform (STFT) is applied with a window size of 1024 frames and a hop size of 320 frames. This process is to obtain spectrograms from the standard time-domain waveforms. Subsequently, Mel filter banks are utilized to the obtained spectrograms. After this, a logarithm operation is performed to derive log Mel spectrograms.

Each of CNN architectures is composed of convolutional layers with a kernel size of 3×3 for CNN10 and CNN14. Batch normalization is applied after

In deep learning research, various studies have adopted transfer learning methodologies, using techniques like embedding extraction and fine-tuning of pre-existing models [13,21], instead of training models from scratch. Both PANNs [11] and DeepSpectrum [1] are highly influential modern libraries designed for audio-based tasks. Among them, PANNs introduce pre-trained audio neural networks for sound event detection. The ability to fix hyperparameters in PANNs provides flexibility to use it as a transfer learning module with pre-existing knowledge. Singh et al. [27] simplified the original PANNs model using a pruning algorithm to remove redundant parameters and reduce computational effort.

To reduce the computational cost, researchers often use pre-trained models with fixed parameters to extract features, and training output layers on the generated embeddings. However, fine-tuning certain layers has been found necessary for specific tasks to achieve excellent performance [11,17]. Earlier layers in convolutional neural networks (CNNs) generally have stronger generalization capabilities than subsequent layers [29], explaining why fine-tuning all layers is essential for achieving good performance. In our study, we aim to explore whether a similar operation is necessary for the model under consideration. We will conduct two experimental designs, freezing the parameters of pre-trained layers or fine-tuning all output layers, to compare the effects of these approaches empirically.

3 Methodology

In our proposed architecture, we have designed two key modules: the pre-trained block and the multi-head attention block. The pre-trained block is a convolutional neural network (CNN) model that we have encapsulated within the PANNs [11] framework. The system's overall structure and the interconnections between these modules are depicted in Fig. 1. In this section, we provide comprehensive explanations of the datasets utilized and the specific application strategies employed for each module.

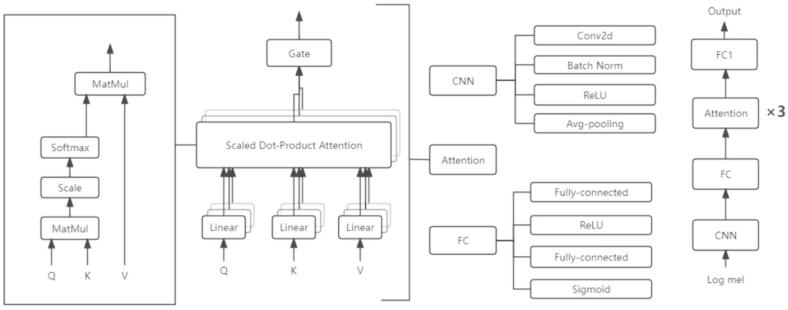


Fig. 1. The structure of proposed multi-head attention block. The number of attention layers will be adjusted to the specific task. 1 and 3 were applied in our experiments.

models, deep learning models do not require handcrafted features but directly learn feature representations from raw data, making them more powerful for processing complex and large-scale datasets and often exhibiting superior performance in specific tasks.

Xu et al. [35] proposed a framework for dual-modal (audio-text) emotion recognition. The framework consists of a parallel convolution module (Pconv) and an attention-based BLSTM [30], with a specific focus on single-modal processing of audio data from the CH-SIMS dataset. By combining Pconv and attention-based BLSTM, the Tensor Fusion Network effectively captures the complementary information from audio and text modalities, enabling more powerful multimodal sentiment analysis. The multiple self-attention mechanism is also a method of sentiment analysis that can enhance modal information [31]. In this paper, we apply transfer learning to a pre-trained CNN model with a multi-head attention mechanism and evaluate the performance of the system in terms of classification accuracy and training time.

2.2 Transformer

The Transformer model possesses several advantages, including its ability to effectively handle long sequences, capture long-range dependencies, and its parallel computing capabilities, making it highly suitable for processing large-scale data. Initially, the transformer model was mainly used in the field of machine translation, but because of its properties, it has gradually been generalized to the field of audio recognition.

In 2015, Chorowski [3] proposed to utilize an attention-based architecture, where the encoder side is a BiRNN structure. This was followed by a study on how transformers can replace RNNs for computation. The combination of CNN and attention mechanism is also a trend in audio emotion recognition, and the self-attention mechanism can express the salient regions of emotion in audio very well [16]. In 2021, Gong et al. Li et al. [15] proposed an Attention pooling method to avoid overfitting of convolutional features input to the fully connected layer. [5]introduced the Audio Spectrogram Transformer (AST), an audio classification model that canceled CNNs. Applying the Transformer encoder output to an audio spectrogram representation. They then proposed a semi-supervised framework [6] that improved the performance of AST by an average of 60.9%.

2.3 Transfer Learning

Transfer learning leveraging knowledge and models learned from one task to improve performance on another related task, reducing the need for extensive training data. It can effectively bypass the time-consuming task of data tagging when discrepancies exist in the feature space or data distribution [25], significantly increasing data mining efficiency. Transfer learning is crucial for multilingual or cross-lingual datasets due to the correlation between languages and speech, enabling the discovery of implicit connections parallelization [28].

collection of data covering Chinese text, images, audio data, and detailed annotations of modality.

In our study, we employed Pretrained Audio Neural Networks (PANNs) that were trained on the comprehensive AudioSet dataset. PANN is a deep learning model architecture crafted for audio data processing, built on the convolutional neural network (CNN) structure. Through fine-tuning on our unique dataset and integrating a multi-head self-attention mechanism, PANNs became more attuned to the specific features of the task and emotional nuances present in speech data, leading to enhanced emotion recognition performance. Our primary contributions include:

- We fine-tuned the pre-trained model on the AudioSet dataset and applied it to CH-SIMS for data preprocessing, yielding results with remarkable generalization capabilities.
- We introduced an architecture that merges CNN with a multi-head attention mechanism, enhancing the model's downstream performance.

2 Related Works

2.1 Speech Emotion Recognition

Over the past nearly three decades, researchers have tried to give machines the ability to understand and express emotions. Currently, the mainstream emotion recognition methods are extracting features that can accurately express emotions and detecting them, either manually or with the help of machines. This field encompasses a wide range of literature and utilizes various English datasets, such as RAVDESS [19], SAVEE [7], and IEMOCAP [2]. AudioSet [4] records a collection of 10-second sound clips including 632 audio event classes and over two million human-tagged clips drawn from YouTube videos. For Chinese language datasets, CH-SIMS [18] is notably prevalent, offering sentiment labels such as Strong Negative, Weak Negative, Neutral, Weak Positive, and Strong Positive. This study contributes to advancing multimodal sentiment analysis and capturing richer representations of sentiment within Chinese language data.

Emotion detection of sound relies on the integration of classical machine learning methods and deep learning techniques. Acoustic features, such as loudness, pitch, and timbre, are extracted and utilized in the algorithm to achieve accurate emotion detection. Spectral features, including Mel Frequency Cepstral Coefficients (MFCC) and their associated features, are also widely used [20]. The demarcation between machine learning and deep learning methodologies primarily resides in their respective approaches to data representations. In machine learning, a set of values is extracted from temporal, frequency, and perceptual domains and then fed into the machine learning algorithm as manually selected or predefined features to establish patterns and relationships for tasks like classification or regression. On the other hand, deep learning employs more complex and elusive algorithms, for example, CNN and attention mechanisms, to automatically learn intricate correlations within data. Unlike the traditional

speeches, such as happiness, sadness, and more. Emotion recognition systems leverage machine learning and deep learning techniques to extract relevant features from speech data, enabling accurate classification of emotions. High-performance SER systems hold significant value across various domains, including human-machine interaction [24], voice assistants [14], and psychological research [8]. They not only help computers better recognize the emotional states of inter-actor, but also pave the way for more personalized and effective human-computer interactions. Advancing SER is one of key objectives in emotion recognition system research. To improve accuracy, researchers employ techniques such as data augmentation and transfer learning, complemented by the use of larger and more diverse speech datasets. These strategies aid in training models proficient at accurately capturing and identifying emotional cues from speech data.

In the task of SER, the objective is to correlate input speech signals with specific emotion categories, thereby determining the underlying expressed emotions. Traditional classification techniques usually rely on probabilistic models, such as the Gaussian mixture model (GMM) [12], hidden Markov model (HMM) [23], and support vector machine (SVM) [26]. However, with the progression of research, various artificial neural network architectures have also been widely utilized, ranging from the simplest multilayer perceptron (MLP) [33], convolutional neural networks (CNNs) [9], to deep architectures like residual neural networks (ResNets) [32] and recurrent neural networks (RNNs) [17] [18]. Particularly, long short-term memory (LSTM)and gated recurrent units (GRU)-based neural networks, which are state-of-the-art solutions in time-sequence modeling, have been ubiquitously applied in speech signal modeling. Additionally, researchers have also proposed various end-to-end architectures aiming to jointly learn both feature extraction and classification [16]. These architectures intensively optimize the identification and association of emotions in speech signals, enhancing the overall performance of SER systems.

SER in Chinese involves identifying and analyzing emotions in Chinese speech data. Chinese-specific speech datasets are used to create diverse databases covering various emotional states. Techniques such as sound signal processing and feature extraction are employed to capture emotion-related features from speech. Machine learning algorithms, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are used for emotion classification. Recent advancements like transfer learning and data augmentation have shown promising results. [34]

The attention mechanism imitates human attention, selectively focusing on different parts of input data and assigning varying levels of importance. Self-attention, used for sequential data, treats each input element as a query, key, and value. Multi-head self-attention extends this concept by introducing multiple attention heads, enabling the model to capture diverse feature representations and enhancing its expressive power.

Our research focuses on Speech Emotion Recognition (SER) in Chinese. We leveraged the CH-SIMS dataset for our study, which provides a comprehensive

Transfer Learning for Audio-Based Speech Emotion Recognition in Chinese: Leveraging Pretrained Models for Improved Performance

Lanke Zhu[1,2], Xinyue Ma[1,2], Rui Zhang[1,2], and Jianbo Zheng[1,2(✉)]

[1] Artificial Intelligence Research Institute, Shenzhen MSU -BIT University, Shenzhen 518172, Guangdong, China
jianbo.zheng@smbu.edu.cn

[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU -BIT University, Shenzhen 518172, Guangdong, China

Abstract. In the field of Speech Emotion Recognition (SER) research, there is a growing emphasis on strengthening model generalization, stepping beyond the traditional classification accuracy metrics. Recent progress in cross-corpus SER has allowed machines to explore relationships among languages from diverse regions. In this paper, we propose an audio emotion recognition model which leverages a pretrained CNN model with a multi-head attention block. To adapt the model for the Chinese dataset CH-SIMS employed in our experiments, we fine-tuned it from a pre-trained English model. The data are categorized into five valence states: negative, weakly negative, neutral, weakly positive and positive. Remarkably, our top-performing model (multi-layer-CNN14) achieves a 24% improvement in accuracy over the baseline. The results highlight the effectiveness of fine-tuning in enhancing speech emotion recognition performance. This study contributes to improving model generalization in transfer learning, nudging us toward a deeper understanding and more accurate recognition of emotions expressed in speech.

Keywords: speech emotion recognition · transfer learning · fine-tuning · attention mechanism · Pretrained audio neural network

1 Introduction

Speech Emotion Recognition(SER) is a vital task in Natural Language Processing (NLP). It aims to detect and recognize the emotions conveyed through

L. Zhu, X. Ma, R. Zhang—These authors contributed equally to this work.

J. Zheng—This work was supported in part by the Shenzhen Sustainable Development Special Project under grant KCXFZ20201221173411032.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 3–14, 2024.
https://doi.org/10.1007/978-3-031-65126-7_1

E-Health Networks I

Autonomous Vehicles