JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 11, pp. 49-56, November 2024
https://doi.org/10.9708/jksci.2024.29.11.049

Analyzing the Impact of Plot Size in Vision-Based Time-Series Sound Classification

Euihyun Jung*

*Professor, Dept. of AI Convergence, Anyang University, Anyang City, Korea

[Abstract]

In recent years, visualizing time-series data as images for use in vision-based Artificial Intelligence (AI) models has gained significant attention. This approach transforms temporal sequences into images that can be processed by deep learning models, such as Convolutional Neural Network (CNN). Although its effectiveness has been demonstrated in various domains, the impact of plot size on model performance remains underexplored. In this study, we investigate the effect of varying plot sizes on classification accuracy by visualizing natural sounds (e.g., cats, crows) and testing five classes of 2,000 samples each using the YOLO model. While training was conducted on 320x320 plots, test sets were generated at six sizes (112x112 to 640x640). Results show that as the plot size of the test dataset diverged from that of the training dataset, both precision and recall decreased, highlighting the importance of plot size consistency in time-series visualization research.

▸ **Key words**: Deep Learning, Convolutional Neural Network(CNN), Time-series, Vision AI

[요 약]

최근 시계열 데이터를 이미지로 시각화하여 영상 인공지능 모델을 활용하는 방법이 주목받고 있다. 이 방법은 시계열 데이터를 이미지로 변환해, 합성곱 신경망(CNN: Convolutional Neural Network)과 같은 딥러닝(Deep Learning) 모델이 처리할 수 있도록 하여, 다양한 분야에서 그 효과가 입증되었지만, 플롯(plot) 크기가 모델 성능에 미치는 영향은 충분히 연구되지 않았다. 본 연구에서는 플롯 크기의 변화가 분류 정확도에 미치는 영향을 조사하기 위해 고양이, 까마귀 등의 자연의 소리를 플롯(plot)으로 시각화하고, 각 2,000개의 샘플로 구성된 5개의 클래스를 YOLO 모델을 통해 테스트하였다. 학습은 320x320 픽셀 크기의 플롯으로 진행되었으며, 테스트 데이터셋(Test dataset)은 112x112에서 640x640까지 6 종류의 픽셀 크기로 생성하였다. 그 결과, 테스트 데이터셋의 플롯 크기가 학습 데이터셋의 플롯 크기와 다를 수록 정밀도와 재현율이 감소하는 것을 확인했으며, 이는 시계열 시각화 연구에서 플롯 크기의 일관성이 중요함을 시사한다.

▸ **주제어**: 딥러닝, 합성곱 신경망, 시계열, 영상 인공지능

• First Author: Euihyun Jung, Corresponding Author: Euihyun Jung
*Euihyun Jung (jung@anyang.ac.kr), Dept. of AI Convergence, Anyang University
• Received: 2024. 10. 04, Revised: 2024. 11. 04, Accepted: 2024. 11. 05.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

Time series data refers to data observed at sequential time intervals. It is widely used across various fields, such as stock prices and weather information. The analysis of time series data commonly employs traditional regression models like ARIMA(Auto-Regressive Integrated Moving Average)[1] as well as deep learning models such as RNN(Recurrent Neural Network)[2]. ARIMA is a statistical model that captures temporal patterns, particularly suited for datasets with trends and seasonality. In contrast, RNN models handle complex, non-linear relationships by leveraging their recurrent structure to remember previous time steps, making them highly effective for non-stationary, multivariate data.

The visualization of time-series data as images for use in Vision AI models, such as Convolutional Neural Network(CNN), has become an innovative approach, attracting growing attention in recent years[3-4]. This method allows for the transformation of complex temporal patterns in time-series data into visual representations (e.g., spectrograms, recurrence plots, Gramian Angular Fields), which can then be processed by image-based deep learning models typically used for tasks such as classification, anomaly detection, or predictive analysis.

One major advantage of this approach is that image-based models are highly effective at identifying spatial features, which can capture underlying patterns in time-series data that might not be as apparent in traditional analysis. For example, spectrograms are frequently used in audio and environmental sound classification as they can capture frequency patterns over time, making them ideal for detecting nuanced sound features in different environments

Studies have demonstrated the effectiveness of this method across various domains. Mushtaq[5] explored environmental sound classification using spectrograms, while Nguyen et al.[6] applied scalograms to heart rate signal analysis, revealing the potential of image-based approaches. Similarly, Barra et al.[7] employed Gramian Angular Fields (GAF) for financial market predictions, and Choi et al.[8] discussed how transforming time-series into images facilitates the use of pre-trained image models for tasks like anomaly detection.

These studies illustrate that transforming time-series data into visual formats is not only a growing trend but also an effective strategy for improving classification performance across various domains. The rationale behind this trend lies in the ability of deep learning models to extract spatial patterns from images that may not be easily discernible through traditional time-series analysis techniques.

However, despite the proliferation of research in this area, there remains a significant gap in the literature regarding the impact of plot size on model performance. While numerous studies have focused on optimizing the visualization techniques and the models used for classification, the question of how varying the size of these plots affects classification accuracy has not been thoroughly investigated. The reason why an in-depth investigation has not yet been conducted is that it is generally accepted that minor variations in the size, position, and orientation of objects in an image do not significantly impact the model's performance[9-10]. This is primarily due to key CNN properties like translation invariance and local receptive fields.

This oversight is crucial, as our study demonstrates that plot size can significantly influence the results, leading to potential misclassification or loss of important features when the test plot sizes differ from those used during training. Thus, our research aims to address this gap, contributing to the understanding of how plot size variability can affect the robustness of vision-based models in time-series classification tasks.

In this study, natural sounds such as those of cats and crows were visualized as time-series plots and classified using vision-based AI. Specifically, we conducted experiments with five classes, each containing 2,000 audio samples, resulting in a total of 10,000 sound data points. These were initially visualized as 320x320 plots and used to train the You Only Look Once (YOLO) model[11]. Subsequently, the classification performance was evaluated by generating plots in different sizes—112x112, 224x224, 320x320, 440x440, 552x552, and 640x640—and comparing the results across these varying sizes.

The performance comparison revealed two key insights. First, as reported in previous studies, visualizing time-series data, including sounds, as plots allows for superior classification performance when using Vision AI. Second, it was observed that when the plot size deviated from the trained 320x320 dimensions—whether by increasing or decreasing—both precision and recall deteriorated. This indicates that analysis through the visualization of time-series data is highly dependent on the plot size used during model training.

The significance of this research lies in the fact that, typically, deep learning models are shared without specifying an exact image size for target classification. However, the results of this study suggest that, in the context of time-series visualizations, it may be necessary to specify the dimensions of the target plots along with model deployment details. This ensures that the model's performance remains consistent and aligns with the intended classification accuracy for different plot sizes.

The structure of this paper is as follows. In Section 2, we review existing studies related to time-series visualization and sound visualization. Section 3 provides an overview of the dataset, its characteristics, and the research methodology. Section 4 discusses the experimental results, and Section 5 presents the conclusions.

II. Related Works

1. Visualization Methods of Time-Series Data

In time series visualization for deep learning applications, various techniques such as Gramian Angular Summation Field (GASF), Gramian Angular Difference Field (GADF), Plot, and Spectrogram are utilized. Each method has its own strengths and weaknesses, making them suitable for different tasks.

GASF is particularly effective at transforming time series data into 2D images, allowing for the capture of overall temporal patterns. This transformation is advantageous as it enables seamless integration with deep learning models designed for image classification. However, one limitation of GASF is that the transformation process may lead to information loss, especially in complex time series data, which can impact the accuracy of the model.

Similarly, GADF excels in capturing local variations and trends within the time series, making it more suitable for analyzing intricate patterns. It also converts time series into 2D images that deep learning models can process. Nonetheless, like GASF, GADF can suffer from information loss during the conversion process, and the resulting representations may be challenging to interpret.

On the other hand, the Plot method offers a straightforward and intuitive way to visualize time series data, providing clear insight into the data's trends and fluctuations. This makes it ideal for exploratory analysis and for researchers who are new to time series visualization. However, its simplicity can also be a drawback, as it may struggle to represent complex patterns in high-dimensional data. Additionally, further preprocessing steps are often required before the data can be used in deep learning models.

The Spectrogram technique provides the unique advantage of analyzing time series data in both the time and frequency domains. This dual

representation is particularly useful for complex signals, such as those found in audio or speech recognition tasks. Spectrograms allow deep learning models to capture both temporal and frequency-based patterns. However, a downside to using spectrograms is that the process of decomposing the time series into frequency components can lead to the loss of high-resolution temporal information, limiting its applicability to only certain types of time series data.

2. Previous Studies

Angelo et al.[12] explores state-of-the-art deep learning architectures like Transformers, specifically addressing the challenges and open issues in time-series forecasting. Li et al.[13] demonstrated that time-series plots could be effectively used for fault diagnosis in mechanical systems, emphasizing the interpretability of visual features. In addition, Sun et al. [14] utilized wavelet transforms to create time-frequency plots, showing improvements in audio signal classification, while Bhowmik, et al.[15] applied time-series heatmaps to detect fraud in financial transactions, illustrating the broader applicability of this approach. Hatami et al.[16] developed a method for converting multidimensional time-series into 2D images for CNNs, setting new benchmarks across various datasets. Braun et al. [17] discusses innovative visualization methods for handling time-series data with varying magnitudes.

Other notable studies include Xie et al.[18] , who implemented Spectral Entropy Maps leading to enhanced anomaly detection. Qin et al.[19] investigated the use of Polar Coordinate Time-Series Transforms (PCTST) for activity recognition in wearable sensor data, establishing a new approach to handling temporal patterns in human activity. Furthermore, Uribarri and Mindlin[20] demonstrated that using Topological Time-Series Embeddings (TTSE) enabled deep learning models to capture long-term temporal dependencies effectively. Dudukcu et al. [21] combines Temporal Convolutional Networks with RNN to predict time-series data.

However, despite the growing body of research in this field, there is still a notable gap in the literature concerning the effect of plot size on model performance.

III. Methodology

1. Data Set

In this study, the ESC-50 dataset[22] from Kaggle was used. The ESC-50 dataset is a labeled collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification. The dataset is widely recognized in academia and has been utilized in over 30 scholarly articles. The study specifically focuses on 5 animal sounds from the dataset including dogs, cats, crows, insects, and pigs. Each class originally contained 40 audio files, with each file lasting 5 seconds. These files were segmented into 1-second intervals, resulting in 200 data points per class.

To increase the dataset size for each class, Gaussian Noise augmentation was applied. Gaussian Noise is a type of noise that follows a normal distribution and is commonly used to augment datasets by introducing slight random variations while preserving the original signal's integrity as formula (1). In the formular, μ is the average value of x and σ is the standard deviation.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (1)$$

By adding Gaussian Noise to the audio data, the dataset was expanded to enhance the model's robustness and ability to generalize to new, unseen data.

When generating Gaussian Noise, the mean is typically set to 0, while the standard deviation is adjusted. The reason for setting the mean to 0 is to ensure that the noise is not biased in one direction, which could alter the overall value of the original

signal. If the mean were not 0, adding the noise would shift the signal's overall amplitude, potentially distorting the original data.

Since the amplitude varies between different audio files, the noise added must be proportional to the original signal to avoid excessive or insufficient noise application. Instead of setting a fixed noise range, such as 255 or -255, which could either overwhelm the signal or add too little variation, the noise range was adjusted using the product of the signal's maximum value and the standard deviation of the noise. This method ensures that the noise is appropriately scaled relative to the signal's amplitude, providing a balanced augmentation effect. By applying Gaussian Noise augmentation in this way, each 1-second audio segment was augmented 10 times, effectively increasing the dataset size by a factor of 10. As a result, each class grew from 200 data points to 2,000 data points, improving the model's ability to handle slight variations and generalize more effectively.

2. Plot Visualization and Model Training

The audio files were converted into time-series plots as shown in Fig. 1, which were then used to train a YOLO model. As shown in Figure 1, each sound such as cat, crow, dog, and insect has distinct characteristics that result in visually distinguishable plot patterns.

The total number of files in the dataset is 10,000, with 2,000 files for each of the 5 classes. For the training, The transformed plots were split into training and testing datasets in a 7:3 ratio.Typically, the training and testing datasets are kept consistent in format when training and evaluating models. However, in this study, to investigate the effect of plot size on model performance, the test dataset was generated in varying sizes: 112x112, 224x224, 320x320, 440x440, 552x552, and 640x640. The trained model was then tested on these different plot sizes.

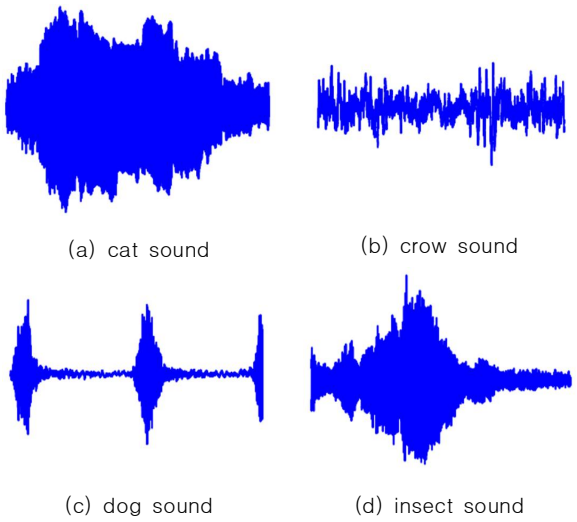

Fig. 1. The sample plots of animal sounds in 1 second

As a result, the 3,000 test files were expanded to a total of 18,000 files for evaluation. The objective was to determine how changes in plot size would affect the model's performance across different classes. The model was trained on a machine with an AMD Ryzen 7 CPU and RTX 2080Ti for a total of 100 epochs over 10 hours.

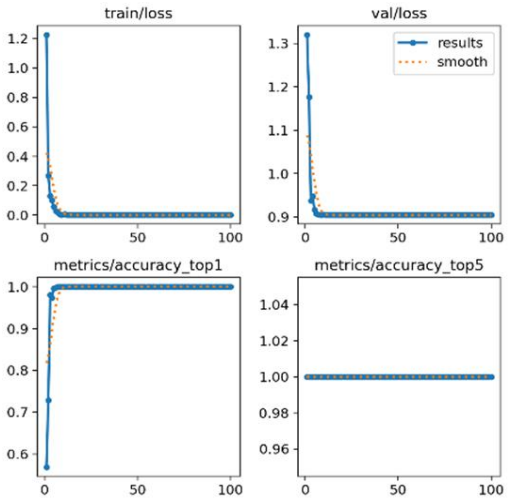

Fig. 2. loss and accuracy of training graphs

As shown in Fig 2, the training graph indicates that both loss and accuracy approach their optimal levels early in the training process. This early convergence suggests that the model was able to quickly recognize the features of the time series images.

IV. Discussions

In the test, the performance of the model was significantly influenced by the size of the plots used during testing. The model achieved perfect accuracy with the 320x320 plot size, identical to the training size. However, performance decreased with both smaller and larger plot sizes. Specifically, the precision, recall, and F1-score metrics dropped notably at the 112x112 and 640x640 sizes, indicating a loss of critical features in smaller plots and the introduction of noise or irrelevant information in larger plots.

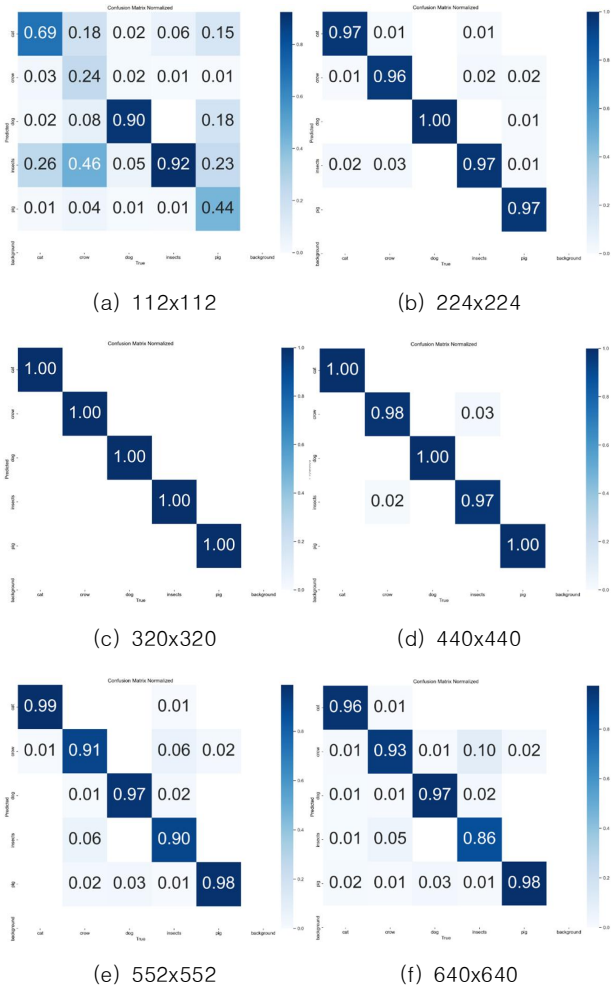

Fig. 3. Confusion Matrix of classification by plot size

Fig. 3 presents the confusion matrices for each plot size. As the test dataset was composed of six different sizes, six corresponding confusion matrices are displayed. As evidenced by the confusion matrices, the accuracy tends to vary more significantly at the extremes of 112x112 and 640x640 plot sizes.

The results confirm our hypothesis that the YOLO model's performance is closely tied to the plot size used during classification. The drop in accuracy with non-matching plot sizes suggests that the model is not invariant to changes in scale, which poses a challenge for applications in environments where plot sizes might vary.

Table 1. Classification performance of the cat sound by plot size

Plot size	Precision	Recall	F1
112x112	0.69	0.55	0.61
224x224	0.97	0.98	0.98
320x320	**1.00**	**1.00**	**1.00**
440x440	0.98	0.98	0.98
552x552	0.95	0.94	0.94
640x640	0.93	0.87	0.90

Table 1 shows the classification performance metrics for the cat sound based on the confusion matrix. As expected, the performance metrics decrease as the plot size deviates from the 320x320 size. Recall and precision tend to be worse, particularly with plot sizes smaller than 320x320. This trend suggests that reduced plot sizes result in diminished information content, which likely impacts the model's classification accuracy.

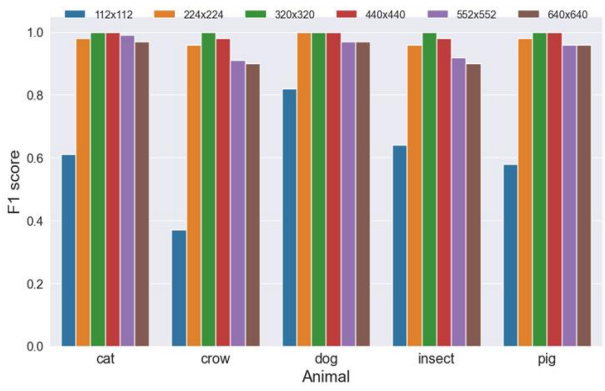

Fig. 4. F1 score of each sound by size

To evaluate the performance across all sound classes, the F1 score was calculated for each class

and plot size, and the results are presented in Fig 4. Consistent with previous findings, the F1 score decreases as the plot size deviates from the 320x320 resolution, regardless of the type of sound.

It means, for time-series visualizations, a mismatch between the plot size used in training and the plot size used during classification can lead to reduced accuracy. Therefore, this study suggests that when publishing models based on time-series visualizations, it is essential to specify the plot size used in model training. This is an important implication for research in time-series visualization.

Although this study identifies valuable insights for time-series visualizations, additional research is necessary to reinforce the theoretical framework. First, since the study focused solely on animal sounds, further investigation on a broader range of sounds is needed to enhance generalizability. Second, it would be beneficial to evaluate Vision AI models beyond YOLO. While YOLO is highly effective, particularly in object detection, it would be insightful to experiment with other models that excel in object classification, such as ResNet. Lastly, further studies should examine the effects of plot resolution and color on model accuracy.

V. Conclusions

In conclusion, this study demonstrates the effectiveness of using time-series visualizations for classifying natural sounds with vision-based AI models. By converting time-series data, such as the sounds of cats into plots, we successfully trained a YOLO model and tested its performance across varying plot sizes. The results confirm the benefits of this approach, as deep learning models like CNN and YOLO can extract spatial patterns from time-series images. However, a key finding from our experiments is the significant impact of plot size on model performance. When the plot size differed from the original 320x320 dimensions used during training, both precision and recall metrics deteriorated. It suggests that visualizing time-series data, discrepancies between the plot size used for training and the size used for classification can result in a decline in accuracy. This study addresses a gap in the current literature by highlighting the importance of plot size in time-series visualization for AI-based classification tasks. Future research should examine how variations in plot characteristics—such as type, resolution, and color—affect model robustness to enhance time-series classification across different domains.

ACKNOWLEDGEMENT

This work was conducted during the sabbatical year.

REFERENCES

[1] R. H. Shumway, D. S. Stoffer, R. H. Shumway, and D. S. Stoffer, "ARIMA models," Time series analysis and its applications: with R examples, pp. 75-163. 2017. DOI:10.1007/978-3-319-52452-8_3

[2] Y. Yu, X. Si, C. Hu, and J. Zhang, J., "A review of recurrent neural networks: LSTM cells and network architectures," Neural computation, Vol. 31, No. 7, pp. 1235-1270, 2019. DOI:10.1162/neco_a_01199

[3] Y. Fang, H. Xu, and J. Jiang, "A survey of time series data visualization research", In IOP Conference Series: Materials Science and Engineering, Vol. 782, No. 2, pp. 022013. IOP Publishing. 2020. DOI:10.1088/1757-899x/782/2/022013

[4] J.F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: a survey," Big Data, Vol. 9. No. 1, pp. 3-21, 2021. DOI:10.1089/big.2020.0159

[5] Z. Mushtaq and S. Shun-Feng, "Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images," Symmetry Vol. 12, No. 11, pp. 1821, 2020. DOI:10.3390/sym12111822

[6] M. T. Nguyen, W. L. Wei and H. H. Jin, "Heart sound classification using deep learning techniques based on log-mel spectrogram," Circuits, Systems, and Signal Processing Vol. 42, No. 1, pp. 344-360, 2023. DOI:10.1007/s00034-022-02124-1

[7] S. Barra, S.M. Carta, A. Corriga, A.S. Podda, and D.R. Recupero, "Deep learning and time series-to-image encoding for financial forecasting," IEEE/CAA Journal of Automatica Sinica, Vol. 7, No. 3, pp. 683-692, 2020. DOI:10.1109/JAS.2020.1003132

[8] K. Choi, J. Yi, C. Park, and S. Yoon, "Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines," IEEE access, Vol. 9, pp. 120043-120065, 2021. DOI:10.1109/ACCESS.2021.3107975

[9] P. Arcaini, A. Bombarda, S. Bonfanti, and A. Gargantini, "Dealing with robustness of convolutional neural networks for image classification," In 2020 IEEE International Conference On Artificial Intelligence Testing (AITest), pp. 7-14. IEEE, 2020. DOI:10.1109/AITEST49225.2020.00009

[10] J. Djolonga, J. Yung, M. Tschannen, R. Romijnders, L. Beyer, A. Kolesnikov, J. Puigcerver, M. Minderer, A. D'Amour, D. Moldovan, and S. Gelly, "On robustness and transferability of convolutional neural networks," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16458-16468. 2021. DOI:10.1109/cvpr46437.2021.01619

[11] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo algorithm developments," Procedia computer science Vol. 199, pp. 1066-1073, 2022. DOI:10.1016/j.procs.2022.01.135

[12] A. Casolaro, V. Capone, G. Iannuzzo, and F. Camastra, "Deep learning for time series forecasting: Advances and open problems," Information Vol. 14, No. 11 pp. 598. 2023. DOI: 10.3390/info14110598

[13] C. Li, J. Xiong, X. Zhu, Q. Zhang, and S. Wang, "Fault diagnosis method based on encoding time series and convolutional neural network," IEEE Access, Vol. 8, pp. 165232-165246, 2020. DOI:10.1109/ACCESS.2020.3021007

[14] X. Sun, P. Liu, Z. He, Y. Han, and B. Su, "Automatic classification of electrocardiogram signals based on transfer learning and continuous wavelet transform," Ecological Informatics, Vol. 69, p. 101628, 2022. DOI:10.1016/j.ecoinf.2022.101628

[15] A. Bhowmik, A., M. Sannigrahi, D. Chowdhury, A.D. Dwivedi, and R.R. Mukkamala, "Dbnex: Deep belief network and explainable ai based financial fraud detection," In 2022 IEEE International Conference on Big Data (Big Data), IEEE, pp. 3033-3042, 2022. DOI:10.1109/BigData55660.2022.10020494

[16] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," In Tenth international conference on machine vision (ICMV 2017), Vol. 10696, pp. 242-249. SPIE. 2018. DOI:10.1117/12.2309486

[17] D. Braun, R. Borgo, M. Sondag, and T. von Landesberger, "Reclaiming the horizon: Novel visualization designs for time-series data with large value ranges," IEEE Transactions on Visualization and Computer Graphics, 2023. DOI:10.1109/TVCG.2023.3326576

[18] W. Xie, Y. Li, J. Lei, J. Yang, J. Li, X. Jia, and Z. Li, " Unsupervised spectral mapping and feature selection for hyperspectral anomaly detection," Neural Networks, Vol. 132, pp. 144-154. 2020. DOI:10.1016/j.neunet.2020.08.010

[19] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.K.R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," Information Fusion, Vol. 53, pp. 80-87. 2020. DOI:10.1016/j.inffus.2019.06.014

[20] G. Uribarri and G.B. Mindlin, "Dynamical time series embeddings in recurrent neural networks," Chaos, Solitons & Fractals, Vol. 154, p. 111612, 2022. DOI:10.1016/j.chaos.2021.111612

[21] H. V. Dudukcu, M. Taskiran, Z. G. C. Taskiran, and T. Yildirim, "Temporal Convolutional Networks with RNN approach for chaotic time series prediction," Applied soft computing, Vol. 133, p. 109945. 2023 DOI:10.1016/j.asoc.2022.109945

[22] K. J. Piczak, "ESC: Dataset for environmental sound classification," In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018. 2015. DOI:10.1145/2733373.2806390

Authors

Euihyun Jung received the B.S., M.S. and Ph.D. degrees in Electronic Engineering from Hanyang University, Korea, in 1992, 1994 and 1999, respectively. Dr. Jung joined the faculty of the Department of Computer Science at Anyang University, Anyang City, Korea, in 2004. He is currently a Professor in the Department of AI Convergence, Anyang University. He is interested in Deep Learning, LLM, and Web3.