CutPaste-Based Anomaly Detection Model using Multi Scale Feature Extraction in Time Series Streaming Data

Byeong-Uk Jeon[1], Kyungyong Chung[2*]
[1]Department of Computer Science, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: jebuk97@kyonggi.ac.kr]
[2]Division of AI Computer Science and Engineering, Kyonggi University
154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, 16227, South Korea
[e-mail: dragonhci@gmail.com]
*Corresponding author: Kyungyong Chung

Received March 18, 2022; revised April 23, 2022; accepted April 25, 2022;
published August 31, 2022

Abstract

The aging society increases emergency situations of the elderly living alone and a variety of social crimes. In order to prevent them, techniques to detect emergency situations through voice are actively researched. This study proposes CutPaste-based anomaly detection model using multi-scale feature extraction in time series streaming data. In the proposed method, an audio file is converted into a spectrogram. In this way, it is possible to use an algorithm for image data, such as CNN. After that, mutli-scale feature extraction is applied. Three images drawn from Adaptive Pooling layer that has different-sized kernels are merged. In consideration of various types of anomaly, including point anomaly, contextual anomaly, and collective anomaly, the limitations of a conventional anomaly model are improved. Finally, CutPaste-based anomaly detection is conducted. Since the model is trained through self-supervised learning, it is possible to detect a diversity of emergency situations as anomaly without labeling. Therefore, the proposed model overcomes the limitations of a conventional model that classifies only labelled emergency situations. Also, the proposed model is evaluated to have better performance than a conventional anomaly detection model.

Keywords: Anomaly Detection, Multi Scale Feature Extraction, Self-Supervised Learning

A preliminary version of this paper was presented at ICONI 2021, and was selected as an outstanding paper.
This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). Additionally this work was supported by Kyonggi University's Graduate Research Assistantship 2022.

http://doi.org/10.3837/tiis.2022.08.018 **ISSN : 1976-7277**

1. Introduction

Currently, as the society has been aging, there have been more emergency situations of the elderly living alone and a variety of social crimes [1][2]. To prevent them, a lot of CCTVs have been installed. However, it is impossible to monitor all CCTVs directly and determine emergency situations. Accordingly, detecting emergency situations through voice is actively researched [3][4]. Cameras are unable to take shots of all areas, so that there are dead zones. Compared to voice equipment, they cost high. Moreover, if cameras are installed in public places, there is concern about invasion of privacy [5][6]. On the contrary, voice equipment costs low and is able to record voices heard from all angles without blind spots. Accordingly, it is necessary to make an anomaly detection model capable of detecting emergency situation through voice. E Principi et al. [7] conduct Voice Activity Detection (VAD) before classifying voice data. In short, they classify only the human voice data detected by the VAD algorithm. Since voice data is classified only with the parts detected by the VAD algorithm, local features are taken into consideration. Therefore, their method fails to consider overall context. Mohammed, M. A. et al. [8] apply pathology detection and classification to voice data with the use of CNN model. They extract features from voice data and classified them, so that accuracy is improved. In their method, features for all voice data are extracted, and then classification is conducted. If one type of CNN kernel is used in classification, global features of voice data only are taken into account. This method fails to consider both global and local features in classification. For this reason, it detects point anomaly as an outlier of a particular point by using a small filter and detects anomaly by using a large filter and checking a change in the global pattern of data. In this way, it takes contextual anomaly and collective anomaly into consideration. As such, it is possible to consider multiple types of anomaly and to make detection. However, it has the limitation of Supervised learning. If a deep learning model has supervised learning, a massive amount of labelled data is needed. In fact, it is hard to secure large labelled data. Moreover, there are many different types of emergencies in real life. Accordingly, if a model performs supervised learning, it detects only emergency situations with labelled learning data, and is highly likely to fail to detect new types of emergency situations. Therefore, it is necessary to design a model that uses self-supervised or unsupervised learning in order to detect a voice in emergency as an anomaly. Since there are various types of emergencies, it is necessary to detect various types of anomaly well in order to detect them as anomaly. To overcome the aforementioned limitations, this study proposes CutPaste-based anomaly detection model using multi-scale feature extraction in time series streaming data. In the proposed method, voice data is converted into spectrogram. From the converted spectrogram image, features are extracted by multi-scale adaptive pooling. At this time, kernels with various sizes are also used to consider both global flow and local features. Accordingly, it is possible to consider various types of anomaly. Finally, the extracted features are put in CutPaste anomaly detection model, and then anomaly detection is conducted. As for model training, normal voice data only are used. The trained model detects a voice in emergency as an anomaly, which has different features from those of normal voice data.

This study is composed of as follows: in chapter 2 is described the frequency analysis through spectrogram conversion and CNN structure and self-supervised anomaly detection; in chapter 3 is described the proposed CutPaste-based anomaly detection model using multi-scale feature extraction in time series streaming data; in chapter 4 is the performance evaluation of the proposed model through the analysis results of the proposed model and objective evaluation; in chapter 5 is drawn the conclusion of this study.

2. Related Works

2.1 Frequency Analysis through Spectrogram Conversion and CNN Structure

Audio data have a variety of components, so that they are expressed in various ways. A way of expression depends on a model. It is possible to improve image classification performance by applying Convolutional Neural Network (CNN) structures, such as VGG [9] and Inception [10]. Accordingly, more research applies a CNN structure to audio data analysis. Wyse, L. et al. [11] proposed the audio-to-spectrogram conversion for a CNN model. For the conversion to spectrogram, Short Time Fourier Transform (STFT) [12] is used for audio signals. The technique is used to cut audio data at a very short time interval (e.g. 0.01 second), and then apply Fourier Transform to each piece. Usually, if L2 normalization is applied, a spectrum is extracted. In spectrogram, both waveform and spectrum feature are expressed. Time and frequency are set as axes, respectively. The intensity of frequency is expressed every frame with the use of concentration or color. By converting audio signal data into 2D image, it is possible to employ a conventional image processing network that consists of three channels. In order for the single channel size value of spectrogram to work in a pre-trained network, it is required to replicate into three channels. The colors of spectrogram are post-processed and synthesized in aesthetic aspect. Even if a spectrogram is generated with Grayscale, it has all related information. In order to put the data in a conventional image processing network, it is necessary to replicate the spectrogram into three channels.

Hershey, S. et al. [13] compares audio classification performance between fully connected model and conventional CNN models. YouTube-100M data sets [11] are used, and audio data are converted into spectrogram. The converted data are put in CNN model. Fully Connected model, and such CNN models as AlexNet [14], VGG [9], Inception V3, and ResNet-50 [15] are compared with each other in terms of classification performance. According to the performance evaluation, CNN models are evaluated to have better classification performance than fully connected network. Therefore, it is wound that CNN models have excellent ability to analyze not only image data but audio data. **Fig. 1** shows data preprocessing and spectrogram conversion process.

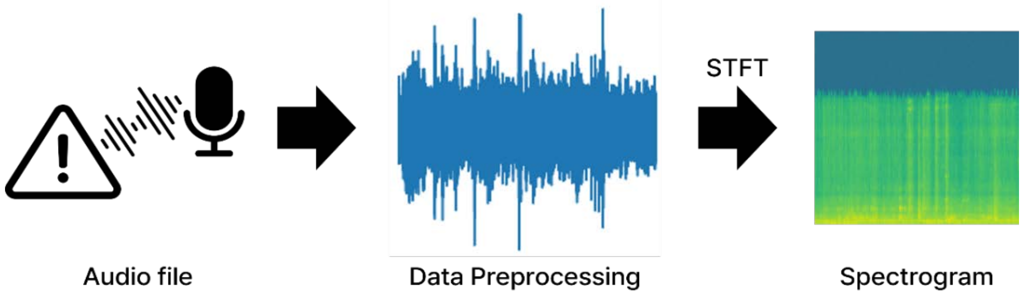
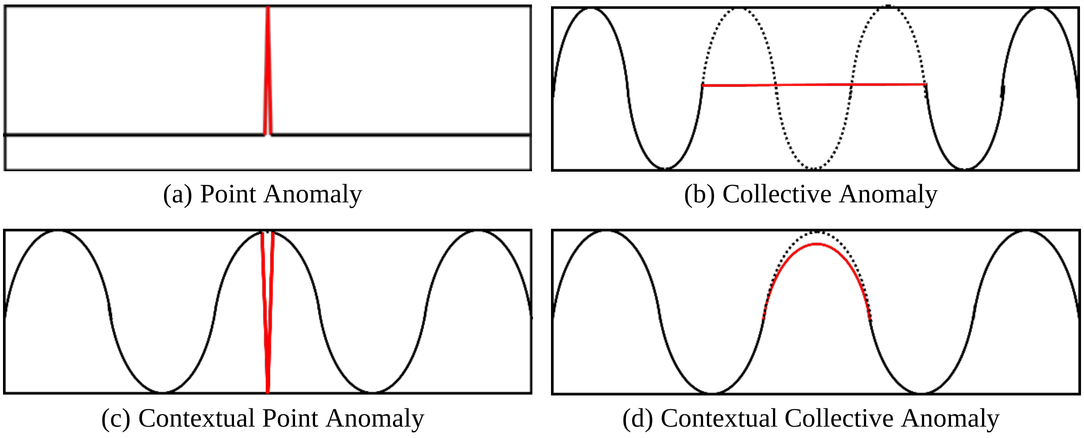


Fig. 1. Data preprocessing and spectrogram conversion process

2.2 Self-Supervised Anomaly Detection

Anomaly detection can be broadly classified into three types: point, contextual, and collective [16]. Point anomaly detection detects anomalies of a specific point. It usually detects outliers that exist in the data. Contextual anomaly detection detects pattern changes in data. Detect outliers that do not follow the overall context. If it is set sensitively, it detects an abnormality even if it is normal, and conversely, if it is insensitive, the abnormality cannot be detected and is missed. Collective anomaly detection detects changes in two or more related data. By comparing two or more features, an anomaly is detected when the change in another feature according to a change in one feature is different from the expected one. **Fig. 2** shows the various types of anomaly.


(a) Point Anomaly

(b) Collective Anomaly

(c) Contextual Point Anomaly

(d) Contextual Collective Anomaly

Fig. 2. Various types of anomaly

An anomaly detection model based on data features analyzes data features first, and then detects the data with different features as anomaly. Models capable of analyzing data features well, such as Auto encoder [17], GAN [18], and CNN models are mainly used, even though they are not for anomaly detection. Therefore, there are anomaly detection models based on GAN, such as AnoGAN [19] and GANomaly [20]. Li, C. L. et al. [21] recently proposed CutPaste, a Self-supervised learning model for anomaly detection and localization. CutPaste utilizes the point that only small parts of an anomaly image are different. By pasting patch, which is a part of an image, to a different random position of the image, the technique augments data in order to use it as an anomaly image in model. The new image obtained by CutPaste is labelled as a virtual anomaly image. With the input data of normal image and the augmented anomaly image, CNN classifier is trained. If a new image is put in the trained CNN classifier, it is possible to verify the feature vector of the data. Gaussian density estimation [22] is applied to the feature vector of the input image, and then the score of anomaly in the unit of pixel is calculated. On the assumption that data follow normal distribution, Gaussian density estimation as an anomaly detection algorithm determines the data located out of the distribution as anomaly data. Since the classification model for images is learned first, it is possible to implement the anomaly detection algorithm more accurately on the basis of the learned image features.

3. CutPaste-based Anomaly Detection Model using Multi Scale Feature Extraction in Time Series Streaming Data

This study proposes CutPaste-based anomaly detection model using multi-scale feature extraction in time series streaming data. The proposed method is capable of detecting voice data in emergency that has different features from those of normal voice data in everyday life. It has three steps. **Fig. 3** shows the process of CutPaste-based anomaly detection model using multi scale feature extraction in time series streaming data.

The first step is spectrogram conversion. As data, the emergency situation voice/sound data set offered by AI Hub is used [23]. This data set includes usual audio data and emergency audio data, each of which has 30 seconds in length. Each data is labelled as to information on the situation at the time of recording. This study views usual voice data as normal data in order for training. For CNN model input, the audio data is applied to STFT algorithm, and is converted into spectrogram. The second step is the feature extraction through multi scale adaptive pooling. Three kinds of pooling layers, each of which has a different size of kernel, are used. Stride or kernel size are adjusted so as to make an output size equal. In this way, the model is able to detect all point anomaly, contextual anomaly, and collective anomaly. The last step is CutPaste-based anomaly detection. For CutPaste model input, three kinds of feature maps extracted from CNN encoder are set to one channel, respectively, and then are merged into the image that has three channel. This image is put in CutPaste-based anomaly detection model, and finally audio anomaly is detected. CutPaste-based anomaly detection model trains only with the feature map converted into normal audio data. Therefore, if anomaly audio data is put in, the model detects different features from normal data used for training, and classifies it as anomaly.

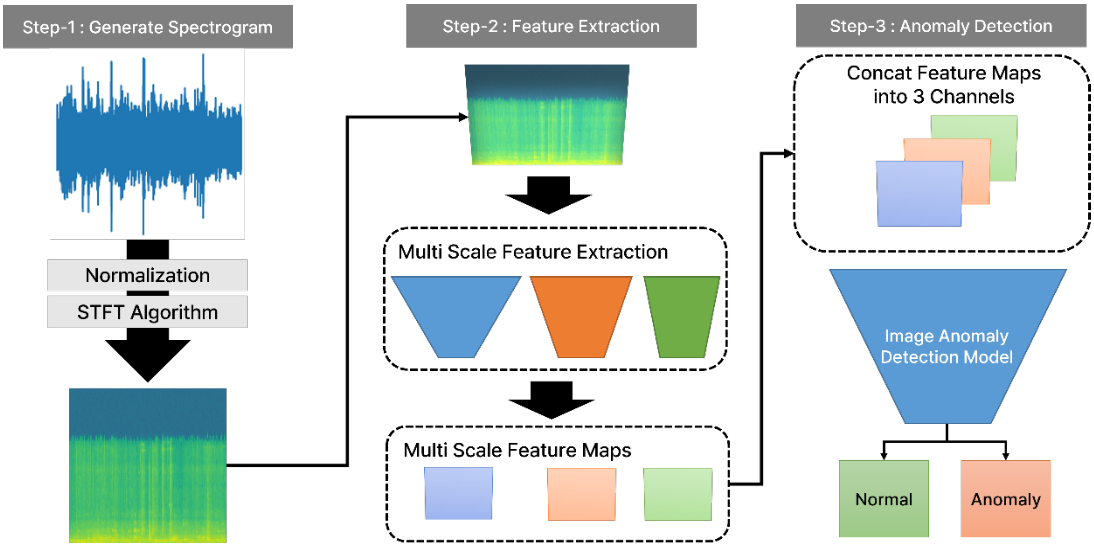
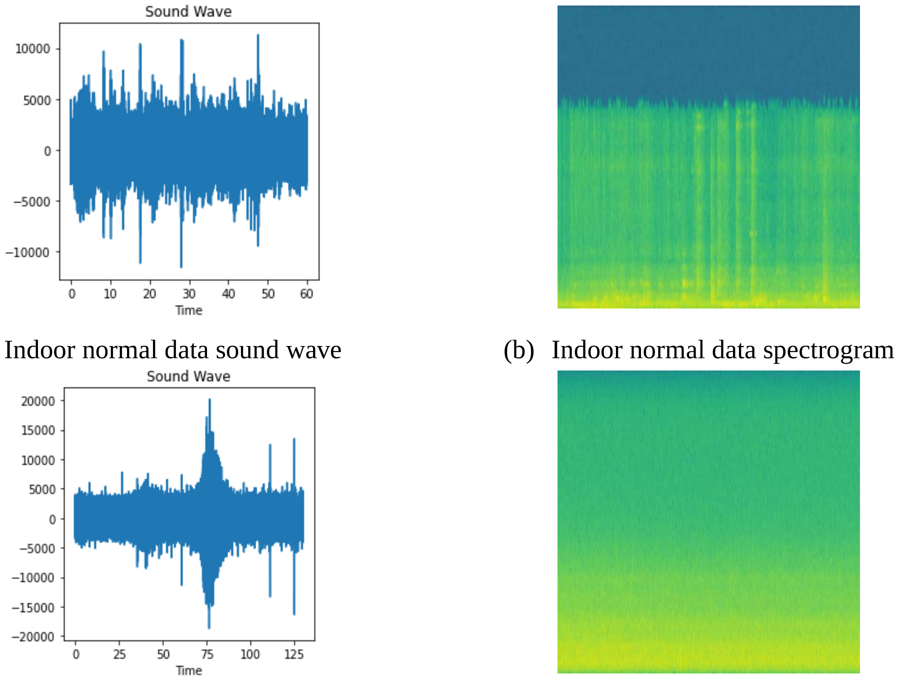

Step-1 : Generate Spectrogram
Normalization
STFT Algorithm
Step-2 : Feature Extraction
Multi Scale Feature Extraction
Multi Scale Feature Maps
Step-3 : Anomaly Detection
Concat Feature Maps into 3 Channels
Image Anomaly Detection Model
Normal
Anomaly

Fig. 3. Process of CutPaste-based anomaly detection model using multi-scale feature extraction in time series streaming data

3.1 Data Collection and Preprocessing

For voice anomaly detection, this study uses the emergency situation voice/sound data set offered by AI Hub. AI Hub is the AI data platform operated by National Information Society Agency in Korea [24]. Thanks to the project of intelligence information industry infrastructure establishment, the AI training data owned by domestic and foreign institutions and companies are open as the data for AI training. As a result, venture enterprises and research institutes are able to receive massive quality data that they hardly secured before. The emergency situation voice/sound data set consists of .wav files that length a total of 3,559 hours. In case of general situations, indoor audio data have 510 hours, and outdoor audio data have 202 hours. In addition, audio data with 12 types of emergency situations including public order security, safety of fire-fighting, natural disaster, accidents and help request, have 2,847 hours. In addition, normal data and emergency situation data are completely labelled. Audio data of normal situations are used as normal data, and audio data of various emergency situations are as anomaly data. A sound is a group of numerous sine waves. Accordingly, it is possible to visualize sine waves of a sound through Fourier transform. If waves are converted into spectrum through Fourier transform, a magnitude of frequency is presented, but the information over time is lost. For this reason, this study sets a short time window with the use of STFT algorithm, and then applies Fourier transform. For visualization, the STFT algorithm based spectrogram conversion method is applied to each audio data. In this way, it is possible to extract spectrogram to visualize the value of the frequency per time in color according to its size. Audio data of general situations length over 30 seconds. In case of audio data of emergency situations, only the audio in an emergency slot is offered. Therefore, audio data of emergency situations have a variety of lengths from 20 to 30 seconds depending on situations. For this reason, in this study, audio data of emergency situations are inserted in the random parts of normal audio data with constant length, and are used as anomaly data. At this time, in order to solve problems, such as a volume difference between audio files, normalization is conducted. In order to make the length of the normal and emergency situation audio data to use constant, the data parts exceeding 30 seconds are cropped, and the data with missing values are removed. Finally, after spectrogram visualization, the image is saved. **Fig. 4** shows the sound wave graphs and spectrogram-converted images of audio data of outdoor and indoor general situations and of emergency situations.

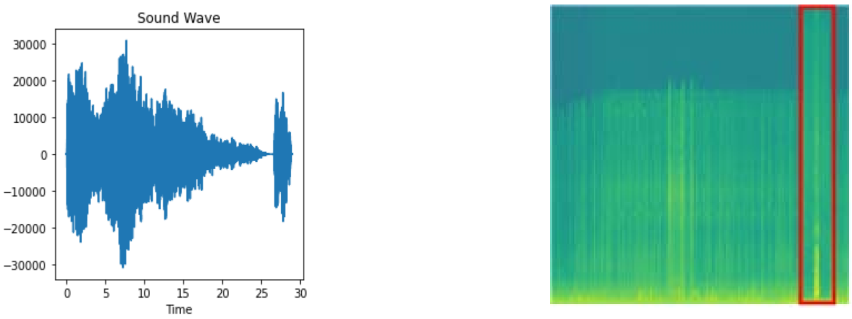


(a) Indoor normal data sound wave

(b) Indoor normal data spectrogram

(c) Outdoor normal data sound wave

(d) Outdoor normal data spectrogram



(e) Anomaly data sound wave

(f) Anomaly data spectrogram

Fig. 4. Sound wave visualization and spectrogram conversion of normal data and anomaly data.

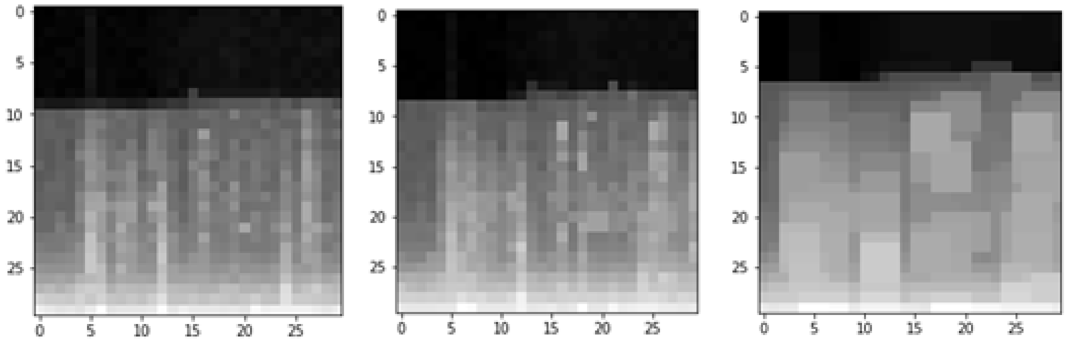
Fig. 4 (a) and (c) show normal data, and **Fig. 4** (b) and (d) present the images after spectrogram conversion. **Fig. 4** (f) shows anomaly data spectrogram. In **Fig. 4** (e), the display part is the anomaly data inserted in normal data. An image size is 900 x 900 pixels. As for normal data, train data 26,622 spectrogram images of train data and 3,997 spectrogram images of validation data, or a total of 30,619 images. As for anomaly data, there are 13,040 spectrogram images of validation data. Because of self-supervised learning, anomaly data do not include train data.

3.2 Feature Extraction using Multi Scale Adaptive Pooling

This study applies adaptive pooling technique to extract multi scale features. Adaptive pooling technique is designed to fix a size of output image regardless of input image sizes [25]. This technique makes it possible to draw output images with an equal size, although kernels have different sizes.

$$stride = \frac{input - kernel}{output - 1} \quad (1)$$

Equation (1) is the equation to adjust a kernel size in consideration of input and output sizes in adaptive pooling. In the equation (1), *input* and *output* are the sizes of input image and output image. *Kernel* is a kernel size of pooling layer. In this study, a fixed output size should be drawn even if a kernel size is different. Therefore, the equation is transformed in order to draw stride values suitable for input, output, and kernel sizes and select and set a value. If the equation (1) is not divisible, it is impossible to guarantee that the output image size of pooling layer is equal to the set size. In order for the equation (1) to be divisible, the stride value and kernel size of pooling layer are set up. To make a kernel size appropriate to an output size, a stride value is fixed. For the equation (1) to be divisible, a number is selected and set up. In this study, an output image size is set as 256 x 256 pixels. Accordingly, the input image with 900 x 900 pixels is resized to that with 768 x 768 pixels. After that, a stride value is fixed to 3, and a kernel size is set to 7, 21, and 63, respectively. **Fig. 5** shows the output result of the multi scale adaptive pooling layer with kernels of 7x7, 21x21, and 63x63 pixel size for one spectrogram image, respectively.



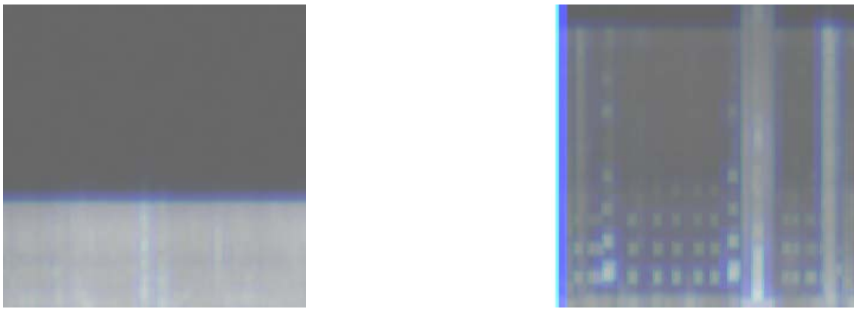
(a) Kernel = 7 (b) Kernel = 21 (c) Kernel = 63

Fig. 5. Features of the speech data with different emotions in the same sentence.

In **Fig. 5** the larger the kernel size, the more global features are derived. Each of the derived features is composed of one channel to compose a merged image of a total of 3 channels. Equation (2) shows the proposed feature extraction process using multi scale adaptive pooling.

$$\begin{pmatrix} \frac{1}{f_h^1 f_w^1} \sum_{u=i'}^{i'+f_h^1-1} \sum_{v=j'}^{j'+f_w^1-1} x_{u,v} \\ \frac{1}{f_h^2 f_w^2} \sum_{u=i'}^{i'+f_h^2-1} \sum_{v=j'}^{j'+f_w^2-1} x_{u,v} \\ \frac{1}{f_h^3 f_w^3} \sum_{u=i'}^{i'+f_h^3-1} \sum_{v=j'}^{j'+f_w^3-1} x_{u,v} \end{pmatrix} with \begin{cases} i' = (i-1) \times stride + 1 \\ j' = (j-1) \times stride + 1 \end{cases} \tag{2}$$

At this time, pooling technique uses average pooling to lessen noise influence. In the equation (2), $z_{i,j}$ means the pixel values of i-column and j-row of merged image. It has the value of 3 channels per pixel. x represents an original image, and i', j' are coordinates of the start point of each pooling kernel. f_k^n represents the row size of nth kernel. f_w^n represents the column size of nth kernel. **Fig. 6** shows the one image generated in the way of merging the outputs of multi scale adaptive pooling.



(a) Merged normal data (b) Merged emergency situation data

Fig. 6. Merged image for the drawn outputs of multi scale adaptive pooling.

In case of merged image, emphasizes areas as shown in **Fig. 6** (a) and (b) are generated. The three images drawn from each scale of pooling layer are merged in one image. Both detailed features and overall features are taken into account. Therefore, it is possible to consider various types of anomaly, including point anomaly, contextual anomaly, and collective anomaly.

3.3 CutPaste-Based Anomaly Detection Model

The merged image drawn from multi scale adaptive pooling is put in CutPaste-based anomaly detection model. If a spectrogram image is put in, a random size and position of area is selected as a patch. At this time, the size of the patch has the min. 0.02 to max. 0.15 ratio to an original image size. The patch of the area is copied and pasted in a random position of the original image. The image augmented in such a way is labelled as anomaly in the model. Fig. 7 shows the structure of CutPaste-based anomaly detection model.

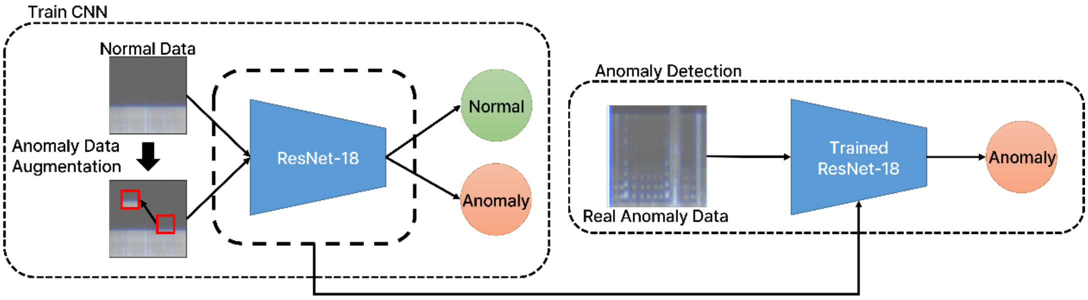


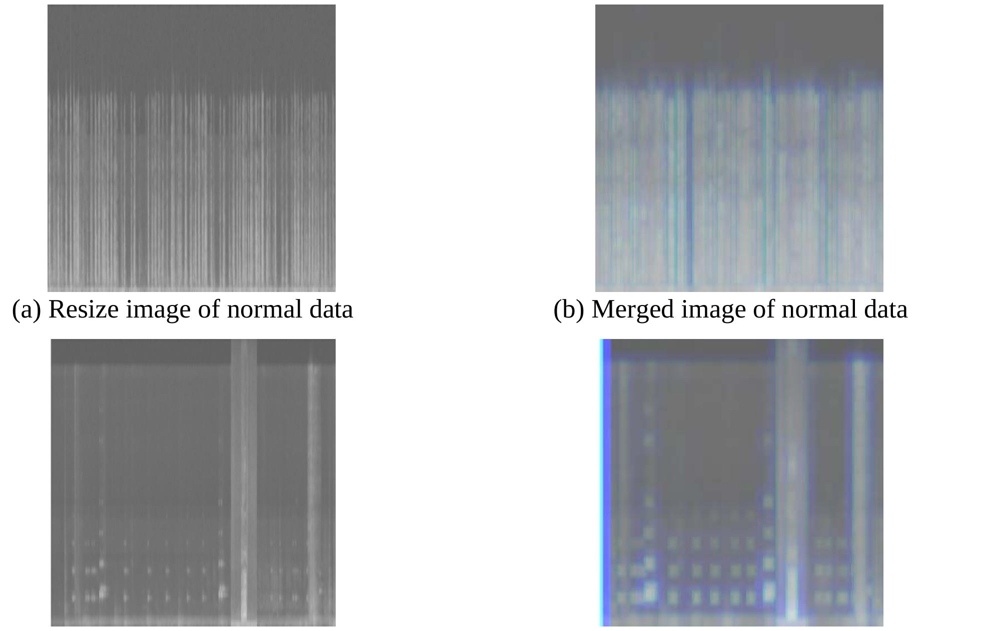
Fig. 7. Structure of CutPaste-based anomaly detection.

In Fig. 7 normal images and the anomaly images augmented by CutPaste technique are finally classified with CNN based binary classification model. In this study, the model is Resnet-18 [15] based binary classification model. The model receives only normal images, and augments and labels anomaly images on the basis of internally input images. With normal images and augmented anomaly images, Resnet-18 model is trained. In the output vector of Resnet-18 model, an anomaly score is finally estimated by Gaussian density estimator. In this way, it is possible to receive normal images only and conduct self-supervised learning for anomaly detection model.

4. Experiments and Results

4.1 CutPaste-Based Anomaly Detection using Multi Scale Feature Extraction

Since vulnerable social groups, including the elderly living alone and children, have poor communication, they request helps by shouting out. The data used in this study are emergency situation voice/sound data. This data set is generated in the way of collecting emergency situation voice/sound data through direct recording and crowd sourcing, and purifying and processing them. It is applicable to respond to emergency situations and welfare safety service for vulnerable classes, including the elderly living alone, children, and women. As for data preprocessing, normalization is applied audio files in order to solve a volume difference. In order to make the length of the normal and emergency situation audio data to use constant, the data parts exceeding 30 seconds are cropped, and the data with missing values are removed. And then, a spectrogram is generated. In this study, voices made in everyday life are defined as normal data, and voice data of emergency situations as anomaly data. Based on them, anomaly is detected. Fig. 8 shows the results from the multi scale adaptive pooling based feature extraction proposed in this study. Unlike a resize image, a merged image has certain areas emphasized.



(a) Resize image of normal data

(b) Merged image of normal data

(c) Resize image of anomaly data

(d) Merged image of emergency situation data

Fig. 8. Results from multi scale adaptive pooling based feature extraction

Fig. 8 (a) and (b) show normal data. **Fig. 8** (c) and (d) present the resize and merged images of anomaly data. In **Fig. 8** (a) and (b), there are less emphasized areas, and the anomaly score difference between CutPaste model and CutPaste using multi scale feature extraction model is small. **Fig. 8** (c) and (d) present data in fire situation. In this case, emphasized areas increased, in comparison to normal data. Accordingly, CutPaste using multi scale feature extraction model drew a higher anomaly score and detected anomaly data more accurately than a conventional CutPaste model. **Table 1** shows the results from the performance comparison between the pooling techniques of the proposed multi-scale feature extraction.

Table 1. Performance comparison between pooling techniques applied to multi feature extraction

Pooling Type	**ROC AUC**
Max Pooling	0.8696602
Average Pooling	**0.9458730**

When average pooing is applied, its ROC AUC is 0.9459, higher (better performance) than ROC AUC (0.8697) of max pooling. That is because max pooling has destructive layers and is greatly influenced by a particular value. If max pooling has an abnormally big noise, the entire kernel area comes out as a value of the noise. Average pooling, however, takes into consideration neighboring data, so that it is not influenced much.

4.2 Experimental Environment and Evaluation Results

The software applications of the experimental environment in the proposed model are Ubuntu 18.04, CUDA version 11.2, and Python 3.7. The hardware consists of NVIDIA RTX 3090 and RAM 24GB. At the time of each model training, batch size is set to 128, and epoch is set to 1000. Multi scale feature extraction and GANomaly model are implemented with TensorFlow 2.4.1. CutPaste model is implemented with PyTorch 1.7.1. A test data set is comprised of 2,000 normal validation data and 2,000 anomaly validation data, which are extracted randomly.

To evaluate the accuracy of anomaly detection, this study applies Receiver Operating Characteristic curve Area Under the Curve (ROC AUC) [26][27]. It represents the width of the bottom area of the ROC curve. The closer the value is to 1, the closer the sensitivity and specificity values are to 1. It means that the related model has good performance. For performance evaluation, GANomaly model without multi scale feature extraction is compared with CutPaste model. Equations (3) and (4) are used to calculate True Positive Rate (TPR) and False Positive Rate (FPR) in order for ROC. In ROC curve, FPR is displayed in x axis, and TPR in y axis. Equation (5) is used to calculate ROC AUC, an index to evaluate model accuracy. The bottom area of ROC curve is used as an index.

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negetive} \tag{3}$$

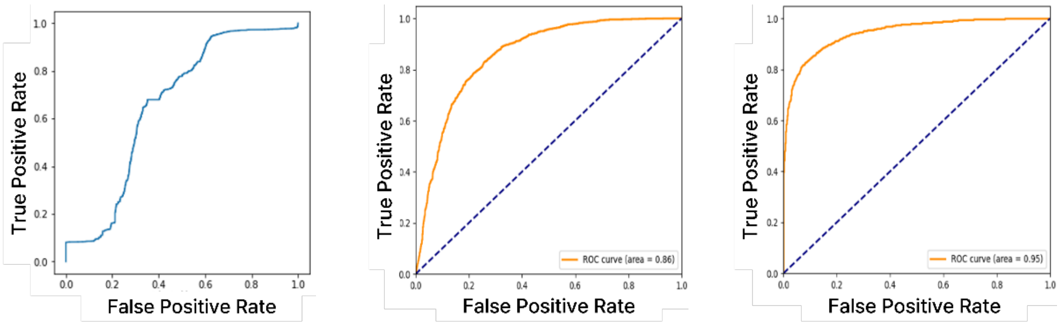
$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive} \tag{4}$$

$$AUC = \int TPR\ d(FPR) \tag{5}$$

In performance evaluation, AUC, which is used mainly as a performance index of classification model, is applied in order to determine if data is normal or anomaly well [28][29]. **Table 2** shows the results of anomaly detection performance comparison between the proposed model, GANomaly model, and CutPaste model. **Fig. 9** shows results of the ROC AUC of each model.

Table 2. Anomaly detection performance comparison between a conventional anomaly detection model and the proposed model

Model	ROC AUC
GANomaly	0.6578606
CutPaste	0.8571284
CutPaste using Multi Scale Feature Extraction (Ours)	**0.9458730**



(a) GANomaly (b) CutPaste (c) Ours

Fig. 9. Results of the ROC AUC

When resize technique was applied simply in line with a model’s input size, GANomaly model’s ROC AUC was about 0.6579, which means bad performance. CutPaste model’s ROC AUC was about 0.8643. As a result, CutPaste model had better performance. According to the performance evaluation of the proposed multi scale feature extraction based CutPaste model, its ROC AUC was about 0.9864, which means excellent performance.

5. Conclusion

These days, deep learning models for detecting emergency situations with the use of videos and voices are actively researched. For emergency detection with the use of videos, cameras have dead zones and cost higher than voice collection equipment. In addition, taking shots of images causes concern for invasion of privacy. Therefore, this study proposes CutPaste-based Anomaly detection model using multi-scale feature extraction in time series streaming data. It extracts a variety of scale features and combines them into one image. In this way, it is possible to consider Point Anomaly, Contextual Anomaly, and Collective Anomaly all at once. According to the performance evaluation of the proposed model, its AUC was about 0.9459 so that the model showed very excellent performance. When resizing, rather than multi scale feature extraction, was applied to input data, the AUC of GANomaly model was 0.6579, and that of CutPaste was about 0.8571. As a result, the proposed model had better performance than other models. With the proposed model, it is possible to detect emergency situations through voice data without blind spots. In addition, since the model is established only with normal data through self-supervised learning and used no labeling, it is possible to detect a variety of emergency situations in real-life. In this study, voice data was used for training and testing of the proposed model. Accordingly, there is a limit that cannot be applied to time series data reflecting spatiotemporal characteristics. In future research, we plan to improve the developing model and evaluate its performance so that it can utilize various time series data.

References

[1] H. Ghayvat, S. Mukhopadhyay, B. Shenjie, A. Chouhan, and W. Chen, "Smart Home Based Ambient Assisted Living: Recognition of Anomaly in the Activity of Daily Living for an Elderly Living Alone," in *Proc. of 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1-5, May. 2018. Article (CrossRef Link)

[2] S. Saqaeeyan and H. Amirkhani, "Anomaly Detection in Smart Homes using Bayesian Networks," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 4, pp. 1796-1816, Apr. 2020. Article (CrossRef Link)

[3] E. L. Piza, B. C. Welsh, D. P. Farrington, and A. L. Thomas, "CCTV Surveillance for Crime Prevention: A 40-year Systematic Review with Meta-analysis," *Criminology & Public Policy*, vol. 18, no. 1, pp. 135-159, Mar. 2019. Article (CrossRef Link)

[4] W. B. Kim and I. Y. Lee, "Secure and Efficient Storage of Video Data in a CCTV Environment," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 13, no. 6, pp. 3238-3257, Jun. 2019. Article (CrossRef Link)

[5] L. Van Zoonen, "Privacy Concerns in Smart Cities," *Government Information Quarterly*, vol. 33, no. 3, pp. 472-480, Jul. 2016. Article (CrossRef Link)

[6] Y. L. Wu, Y. H. Tao, and C. J. Chang, "A Comparative Review on Privacy Concerns and Safety Demands of Closed-Circuit Television among Taiwan, Japan, and the United Kingdom," *Journal of Information and Optimization Sciences*, vol. 38, no. 1, pp. 173-196, Feb. 2017. Article (CrossRef Link)

[7] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An Integrated System for Voice Command Recognition and Emergency Detection Based on audio signals," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5668-5683, August. 2015. Article (CrossRef Link)

[8] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. K. A. Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. Al-Dhief, "Voice Pathology Detection and Classification using Convolutional Neural Network Model," *Applied Sciences*, vol. 10, no. 11, p. 3723, May. 2020. Article (CrossRef Link)

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, Sep. 2014. Article (CrossRef Link)

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826, 2016. Article (CrossRef Link)

[11] L. Wyse, "Audio Spectrogram Representations for Processing with Convolutional Neural Networks," *arXiv preprint arXiv:1706.09559*, 2017. Article (CrossRef Link)

[12] H. J. Kwon, M. J. Kim, J. W. Baek, and K. Chung, "Voice Frequency Synthesis using VAW-GAN based Amplitude Scaling for Emotion Transformation," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 16, no. 2, pp. 713-725, Mar. 2022. Article (CrossRef Link)

[13] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (icassp)*, pp. 131-135, 2017. Article (CrossRef Link)

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, Jun. 2017. Article (CrossRef Link)

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016. Article (CrossRef Link)

[16] M. Ahmed, A. N. Mahmood, and J. Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19-31, Jan. 2016. Article (CrossRef Link)

[17] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Autoencoder-Based Network Anomaly Detection," in *Proc. of 2018 Wireless Telecommunications Symposium (WTS)*, pp. 1-5, Apr. 2018. Article (CrossRef Link)

[18] H. J. Kwon, D. H Shin, and K. Chung, "PGGAN-Based Anomaly Classification on Chest X-Ray Using Weighted Multi-Scale Similarity," *IEEE Access*, vol. 9, no. 1, pp. 113315-113325, Aug. 2017. Article (CrossRef Link)

[19] D. H. Shin, R. C. Park, and K. Chung, "Decision Boundary-Based Anomaly Detection Model using Improved AnoGAN from ECG Data," *IEEE Access*, vol. 8, no. 1, pp. 108664-108674, Jun. 2020. Article (CrossRef Link)

[20] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-Supervised Anomaly Detection via Adversarial Training," in *Proc. of Asian Conference on Computer Vision*, pp. 622-637, May. 2019. Article (CrossRef Link)

[21] C. L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-Supervised Learning for Anomaly Detection and Localization," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664-9674, 2021. Article (CrossRef Link)

[22] X. Shi, Q. Kang, M. Zhou, A. Abusorrah, and J. An, "Soft Sensing of Nonlinear and Multimode Processes Based on Semi-Supervised Weighted Gaussian Regression," *IEEE Sensors Journal*, vol. 20, no. 21 pp. 12950-12960, 2020. Article (CrossRef Link)

[23] Emergency Voice/Sound, AI hub, Last modified on: Nov 18. 2021. [Online]. Available: Mar 15. 2022, https://aihub.or.kr/node/30742.

[24] AI Hub, Last modified on: Nov 18. 2021. [Online]. Available: Apr 13. 2022, https://aihub.or.kr.

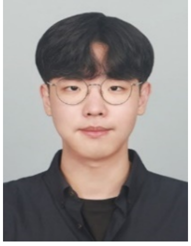
[25] B. McFee, J. Salamon, and J. P. Bello, "Adaptive Pooling Operators for Weakly Labeled Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180-2193, Aug. 2018. Article (CrossRef Link)

[26] S. Naseer and Y. Saleem, "Enhanced Network Intrusion Detection using Deep Convolutional Neural Networks," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 10, pp. 5159-5178, Oct. 2018. Article (CrossRef Link)

[27] J. W. Baek and K. Chung, "Pothole Classification Model Using Edge Detection in Road Image," *Applied Sciences*, vol. 10 no. 19, pp. 6662-6680, Sep. 2020. Article (CrossRef Link)

[28] H. Yoo, R. C. Park, K. Chung, "IoT-based Health Big-data Process Technologies: A Survey," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 15, no. 3, pp. 974-992, Mar. 2021. Article (CrossRef Link)

[29] H. Yoo, K. Chung, "Classification of Multi-Frame Human Motion Using CNN-based Skeleton Extraction," *Intelligent Automation & Soft Computing*, vol. 34, no. 1, pp. 1-13, Apr. 2022. Article (CrossRef Link)



Byeong-Uk Jeon received his B.S. degree from the Division of Computer Science and Engineering, Kyonggi University, South Korea, in 2022. He is currently in the Master course of Department of Computer Science, Kyonggi University, Suwon, South Korea. He has worked as a researcher at the Data Mining Lab., Kyonggi University. His research interests include data mining, big data, deep learning, machine learning and computer vision.

Kyungyong Chung received his B.S., M.S., and Ph.D. degrees in 2000, 2002, and 2005, respectively, from the Department of Computer Information Engineering, Inha University, South Korea. He has worked for the software technology leading department of the Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a professor at the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he has been a professor in the Division of AI Computer Science and Engineering, Kyonggi University, South Korea. He was named in 2017 as a Highly Cited Researcher by Clarivate Analytics. His research interests include data mining, artificial intelligence, healthcare, knowledge systems, HCI, and recommendation systems.