International Journal of Applied Engineering Research ISSN 0973-4562 Volume 17, Number 1 (2022) pp. 30-35
© Research India Publications. http://www.ripublication.com

Deep Learning Based Infant Cry Analysis Utilizing Computer Vision

Joohyun Cha[1,*], Gimin Bae[2]

[1]*Department of Computer Science, University of Illinois Urbana Champaign, Illinois, USA.*

[2]*Department of Computer Engineering, Hanyang University, Seoul, Republic of Korea.*

**(Corresponding Author's e-mail: joohyun6@illinois.edu)*

Abstract

Effort to understand infants' cry is crucial as crying is the main form of communication for infants and sensitively responding to their cry is closely related to his/her development. This paper proposes a method that classifies a video of a crying infant using deep learning techniques to effectively accomplish this task. Specifically, the method utilizes various audio feature extraction techniques (Mel Frequency Cepstral Coefficient and Short-Time Fourier Transfrom), and classification models such as autoencoder, deep residual network, and concatenate layer. The experiment results show that the proposed model obtains high accuracy in interpreting infants' cry compared to other machine learning based models.

Keywords: Infant cry, Feature extraction, Deep learning, Autoencoder, ResNet, Ensemble model, Rolling Prediction Average

1. INTRODUCTION

Infancy is a period in which important, fundamental processes such as cognitive, social, and emotional development takes place. Since proper emotional development plays a crucial role in the success and happiness in their later lives, it is necessary to look at the factors that bring about the negative emotional development of infants.

One of the main factors mentioned in relation to the causes of this negative emotional development is parenting. Parenting is an important factor that affects the most basic elements of an infant including the child's emotions. Thus, in previous studies, parent care is considered an independent variable, and studies refer to this variable as the parent effect. According to a research, infants in families whose parents do not show a loving attitude or are not sensitive to a child's specific behavior are more likely to develop negative emotions than those in families that do not [1]. The research especially focused on the relationship between sensitive reactions and negative emotions of infants, and infants with mothers who react more sensitively to them went through a more positive emotional development compared to the infants with mothers who did not. Based on these results, the necessity to establish a way of communication between infant and parents.

Recent advancement in technology has provided us a way to accomplish this task. There is active research on signal and image processing teqhniques using machine learning and deep learning. Building up on these techniques, video understanding has become a popular research direction in the field of computer vision [2], [3], [4], [5].

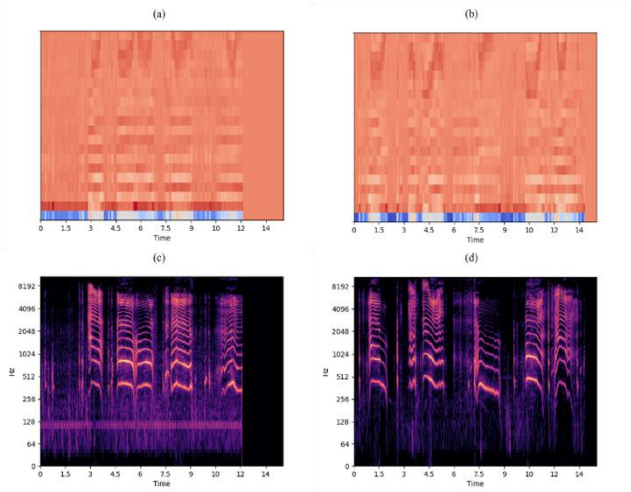

Figure 1. MFCC and STFT spectrogram

This paper presents a deep learning-based video classification method that finds out what infants want by their crying sounds and facial expressions. For the steps, we first split the video of crying infant into image and audio data. Then for audio data, we go through a feature extraction process, then classify it using a Autoencoder + Deep Neural Network (DNN) model. For image data, we use ResNet-50 model to classify each frame and then use a technique called rolling prediction average to produce a label for the input video as a whole. Finally, we use and ensemble model to combine the results and produce the final output. We expect our method to provide help in understanding cry of infants which can ultimately help lessen the burden of parenting.

2. RELATED WORKS

Infants' crying is a basic language for instinctively expressing their abnormal conditions, such as hunger, pain, sleepiness and discomfort. Studies of infants' cry has been conducted since the 1940s. For example, one study tried to analyze the crying characteristics of infants and children to diagnose diseases [6]. This constant interest in the crying of infants led to attempts to categorize the crying according to the reason why the infants are crying. By various research, it has been revealed that crying can be classified into various types, and studies continue to classify crying in various ways [7], [8].

Especially, recent advances in machine learning, deep learning, and signal processing technologies have enabled faster and more accurate analysis than previous studies that relied simply on expert analysis [9]. These developments have enabled crying analysis to classify not only serious situations like pain and disease, but also to find out infant's discomforting situation (sleep, hunger, etc.). As such, many studies have analyzed infant crying to use various techniques in the field of speech recognition and analysis; one research, for example, used Mel Frequency Cepstral Coefficient (MFCC) to extract the characteristics of infant crying, and analyzed the characteristics using Time Delay-Natural Network (TDN) to distinguish children from hearing disorders and suffocation [10]. Another study proposed a method to utilize Support Vector Machine (SVM) to analyze crying sounds with features extracted through Fast Fourier Transform (FFT) [11]. In addition, other studies divided the data into five types of crying: hunger, discomfort, burp, pain, and sleep, and classified specific crying into appropriate data crying types by characteristics extracted through MFCC or PLP using several machine learning techniques such as SVM and Random Forest [12], [13].

However, the past research shows some possible improvements to be made. First, no study utilized video data and only used the audio data. Since facial expressions are proved to be an important means of communication for infants, this paper uses video data, which includes the image data which contains the facial information as well as the audio data that is conventionally used in analyzing infants' cries. Second, the previous research selected only one method of feature extraction to train in the machine learning or deep learning model. These factors may lead to biased results; thus, this paper proposes a new method to classify video of crying infants by using more audio features and image data to obtain a more reliable result.

Table 1. Audio feature credibility test data

Category	Original Data #	Test Data #
Pain	38	20
Stomachache	16	16
Discomfort	27	20
Fatigue	24	20

Table 2. Audio feature credibility test result

Classifier	STFT Accuracy	MFCC Accuracy
SVM	45.83%	66.67%
Naïve Bayes	75.83%	77.50%
Random Forest	79.17%	83.33%
Decision Tree	79.17%	79.17%
K-NN	54.17%	50.00%
DNN	85.59%	84.49%

3. METHOD

3.1. Dataset

The dataset we use for this experiment consists of 800 15-second videos of crying infants, annotated with three class labels – sleepy, hungry, or want to be hugged. The three class labels were selected since it is known to be the three state that is hardest to differentiate each other from. The label of each video was determined by what the parent had to do to make the infant to stop crying. We ensure that all videos have the face of the infant in the center and facing the forward direction. From the video dataset, we use OpenCV to extract frames in the rate of 30 fps and create a dataset that contains 360,000 images of infant faces.

3.2. Audio Data Preprocessing

We first extract the audio data from the original video data via ffmpeg. Then, since the audio signal values from one data to another, we first normalize the values using the Standard Scaler and then use the Min-max Scaler to limit the range of the values from 0 to 1.

We then employ two major audio feature extraction techniques that is proven to be efficient in previous studies [14], [15], namely Mel-Frequency Cepstral Coefficients (MFCC) and Short-Time Fourier Transfrom (STFT). Each feature provides benefits in the classification task. MFCC acts as a backbone by producing a robust feature when the signal is affected by noise, thus reducing error [16]. STFT provides an advantage by incorporating temproal information and reavealing frequency contents of the data at each time point in the signal [17].

Figure 1 is an example of the spectrogram obtained through the MFCC and STFT methods. (a) is the Spectrogram in hungry state, (b) the MFCC Spectrogram in sleepy state; (c) is the STFT Spectrogram in the same state as (a); and (d) is the STFT Spectrogram in the same state as (b).

3.3. Audio Feature Validation

To verify the validity of sound characteristics extracted by the proposed STFT and MFCC methods, cry classification is tested using Donateacry-corpus data. Donateacry-corpus is a collection of infant crying sounds built through a cry campaign, and there are infant crying data collected from Android and iOS. The corresponding crying data can be distinguished by the sex of an infant (male, female), age (0 - 4 weeks, 4 - 8 weeks, 2 - 6 months, 2 months - 2 years, or more), type of crying (e.g., need for abdominal pain, discomfort, fatigue, loneliness, cold or hotness, surprise, unknown).

Table 1 shows the number of data for each category in the experiment. Tests to verify the validity of the speech characteristics use data of four types of crying (e.g., pain, stomachache, discomfort, fatigue).

Table 2 shows that in the test, SVM, Naive Bayes, Decision Tree, Random Forest, K-NNN, and DNN algorithms were used, and 70% of the data were used for training and 30% of the data were used to test. The experiment result confirmed that

both MFCC and STFT obtained highest accuracy in Deep Neural Network and could be used for classification.

3.4. Deep learning algorithms

3.4.1. Autoencoder

An autoencoder is a type of neural network that is designed to encode the input into a compressed and meaningful representation, and then decode it back to a form that is similar to the input. The following equation shows the task of a typical autoencoder.

$$\underset{A,B}{\operatorname{argmin}}\, E[\Delta(x, B \cdot A(x))] \quad (1)$$

The task is to learn functions $A : \mathbb{R}^n \rightarrow \mathbb{R}^p$ (encoder) and $B : \mathbb{R}^p \rightarrow \mathbb{R}^n$ (decoder) that satisfies the equation above where E is the expectation over the distribution of x, and Δ is the reconstruction loss function that measures the distance between the output of the decoder and the input.

Autoencoders are used for various purpose including denoising [18]. Denoising autoencoders remove the noise that disrupts meaningful data and reconstructs a clean version of the input. Related work shows that autoencoders are very effective compared to other models in this specific task [19].

3.4.2. ResNet

He et al. proposed the residual network (ResNet) to learn residual functions with reference to the layer inputs, instead of learning unreferenced functions Instead of hoping each few stacked layers directly fit a desired underlying mapping, ResNets let these layers fit a residual mapping [20]. For the desired underlying mapping $H(x)$, we let the stacked nonlinear layers fit another mapping of

$$F(x) := H(x) - x \quad (2)$$

Then, the originl mapping is recasted into $F(x) + x$. This allows the model to react more sensitively to the change of input data, which is crucial for image classification. Accordingly, ResNet has shown state-of-the-art results in image classification [21], thus we choose to use this model for our task.

3.5. Proposed Model

This paper proposes a deep learning-based ensemble model that receives three forms of input: MFCC feature, STFT feature, and face images. The model then classifies the input data using different submodels depending on the type of input. Finally, the model combines the results of the submodels to produce a single output. Figure 2 shows the diagram of the proposed model.

To classify the audio features, we use a model that combines the Autoencoder and Deep Neual Network (DNN). As shown in Figure 2, we use only the front part of the Autoencoder to remove the noise. Then starting from the bottleneck, we use a DNN model to perform thae actual classification. For the optimizer, we use the root mean squared propagation (RMSProp).

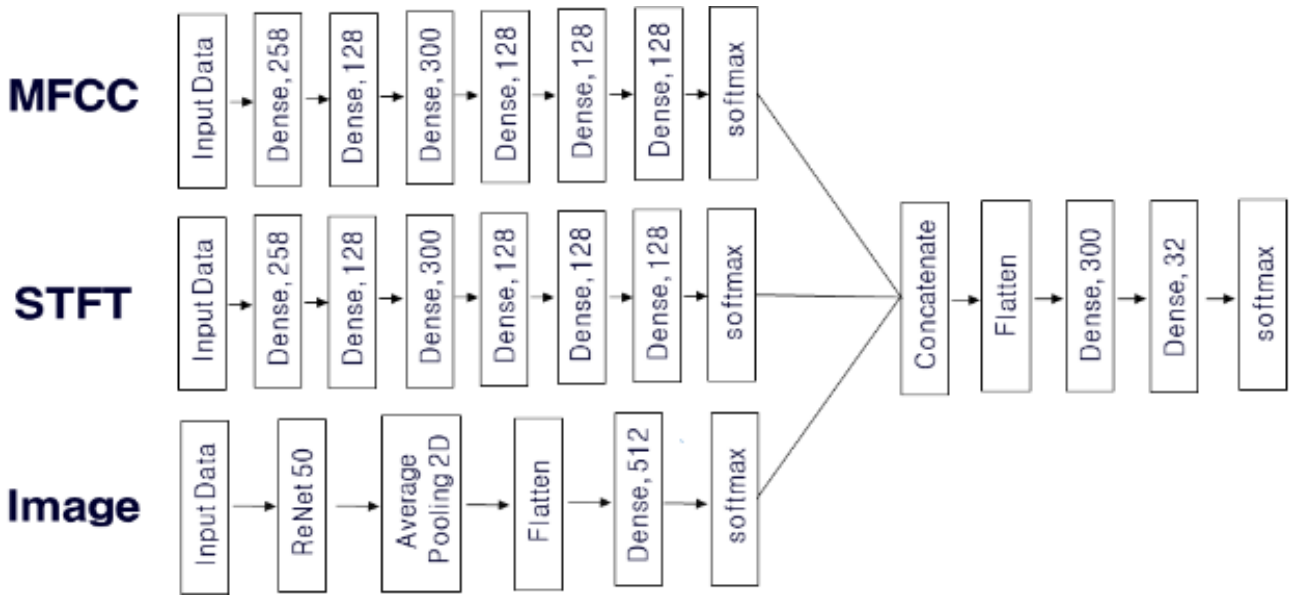

Figure 2. Proposed Model

To classify the images, we use the ResNet-50 model. We use weights of ImageNet and freeze the model so that the model is not trainable. We then construct a second model consisting of an input layer that takes the output of the ResNet-50 model as input, a pooling layer, a flatten layer, a dense layer, and the output layer that produces the result using the softmax function as an activation function. Our final model is a combination of the two models where only the second model can be trained. We choose to freeze the first model since we observed that it empirically produces a better result than allowing to train both models.

For video prediction, we use a technique called rolling prediction averaging [22]. Since a video consists of numerous frames and we make predictions for individual frames, if the frames inside one video have different results, 'prediction flickering' – a stuation where the label of the video constantly change – occurs. To prevent this, we maintain a deque that contains value of past predictions. We use this deque to compute the average of the predictions and choose the label with the largest corresponding probability.

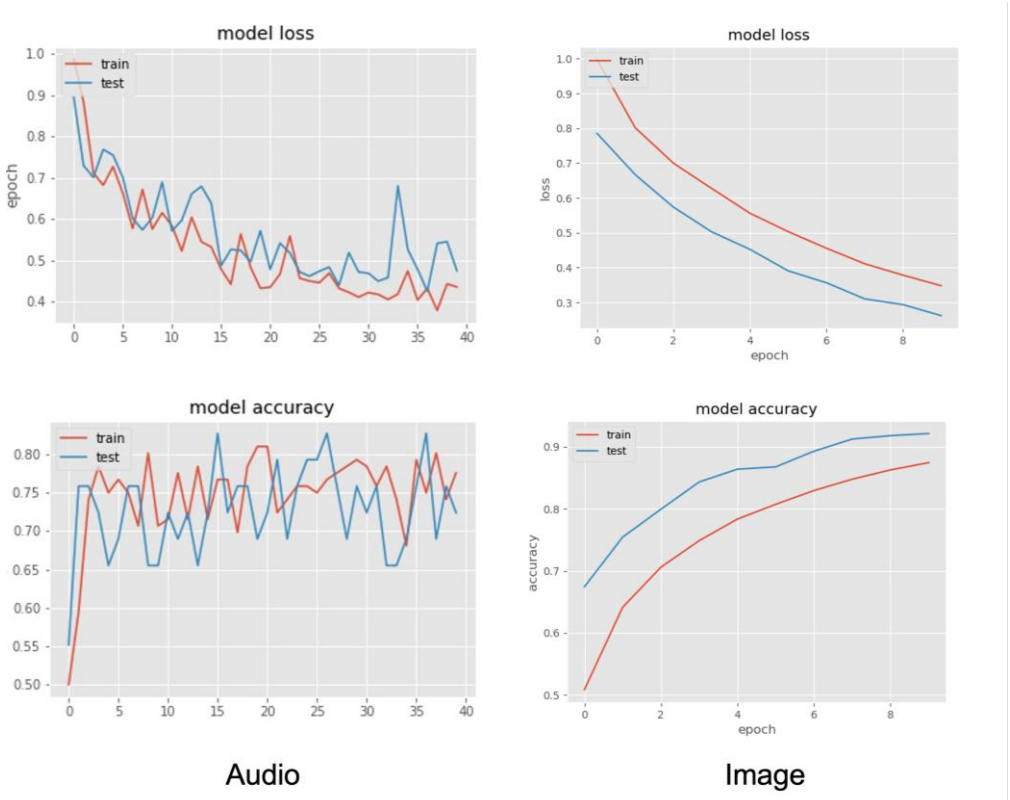

Figure 3. Audio/Image Model Loss and Accuracy

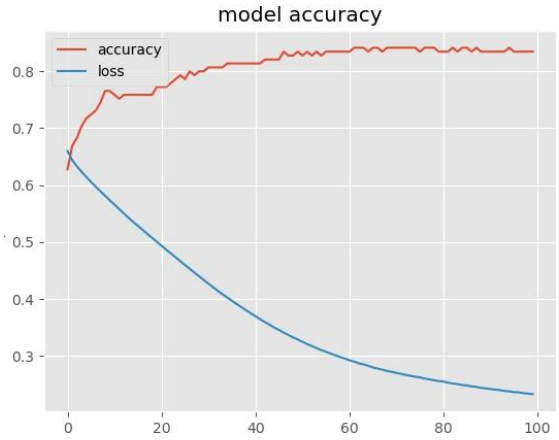

Figure 4. Final Model Loss and Accuracy

Our last model is the concatenate model that combines the outout of each of the three inputs t produce a single output. We choose this method because ensemble models empirically demonstrate to produce reliable results and show higher accuracies compared to single models, [23], [24]. Also, since this experiment manages diverse types of inputs, ensemble models area good fit. This model uses four layers to complete the task. The first layer is the concatenate layer that takes the output of previous models as input and combines them. Then, the flatten layer collapses the spatial dimensions into the channel dimension. After that, we add a dense layer using the rectified linear activation function. Finally, the output layer with the softmax activation function produces the final classification result.

4. RESULTS AND DISCUSSION

In this paper, we use 80% of the data for training and 20% for testing. We chose the number of epochs when the accuracy stopped improving on the test sets. Training for image classification model usually stops after 10 epochs, training for audio models stop at after about 40 epochs, and training for concatenate model (final model) stops after 100 epochs. Figure 3 shows the loss plot per epoch for the audio and image model. Figure 4 shows the loss and accuracy plot per epoch for the final model. The steady increase in accuracy and decrease in loss shown in the graph confirms that the model is training well.

Table 3. Audio feature credibility test data

Category	SVM	AE+DNN	Proposed Model
Sleepy & Hungry	80.92%	97.14%	97.14%
Sleepy & Hugging	68.42%	100%	100%
Hugging & Hungry	66.66%	80%	86.67%
Total (Sleepy & Hungry & Hugging)	65.51%	93.75%	97.41%

To be objective, we compare our proposed model to two different models using the same dataset in the same conditions. The first model is a typical support vector machine (SVM) model which is a commonly used machine learning method. The second model is a deep learning model that combines all the three inputs used in this paper and classifies with a single autoencoder + DNN model. As shown in Table 3, our proposed model has the highest accuracy of 97.41%.

Also, by checking the accuracy of two category classifications, it could be seen that hungry and want to be hugged is more difficult to differentiate than sleepy, hungry, and want to be hugged. This showed that features of hungry and want to be hugged are similar, but still more data will be needed to confirm this fact.

5. CONCLUSION

This paper proposes a methodology of classifying cries of infants. The audio and image data were first extracted from the original video data and then went through a preprocessing stage to produce three types of input: MFCC, STFT, and image data. Finally, our proposed model classifies the input using different submodels for different types of data and combining their output.

By comparing our model to other deep learning algorithms, we see that our proposed model is strong compared to other models in terms of accuracy. Our model showed 97.41% accuracy which is 3.56% higher than that of the autoencoder + DNN model, and 31.90%. higher than that of the SVM model. This proves that the methodology this paper proposes establishes a better way of communication between infants and people who take care of them by interpreting the cries of infants more accurately.

REFERENCES

[1] M. Fish *et al.*, "Patterns of mother-infant interaction and attachment: A cluster-analytic approach," *Infant Behavior And Development*, vol. 18, pp. 435–446, 1995

[2] J. Sivic and A. Zisserman., "Video Google: A text retrieval approach to object matching in videos," *ICCV*, 2003

[3] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid *et al.*, "Evaluation of local spatio-temporal features for action recognition," *BMVC*, 2009

[4] J. Liu, J. Luo, M. Shah, "Recognizing realistic actions from videos," *CVPR*, 2009

[5] J. C. Niebles, C. W. Chen, L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," *ECCV*, pp. 392–405, Springer, 2010

[6] O. C. Irwin, "Vowel elements in the crying vocalization of infants under ten days of age," *Child Development*, pp. 99–109, 1941

[7] J S. Bano, K. M. Ravikumar, " Decoding Baby Talk: Basic Approach for Normal Classification of Infant Cry Signal," *International Journal of Computer Applications*, 2015

[8] S. Bano, K. M. Ravikumar, " Decoding Baby Talk: Basic Approach for Normal Classification of Infant Cry Signal," *International Journal of Computer Applications*, 2015

[9] O. Aomar *et al.*, "Machine Learning Approach for Infant Cry," *International Conference on Tools with Artificial Intelligence*, 2017

[10] J. Hartarto *et al.*, "Infant's Cry Sound Classification using MelFrequency Cepstrum Coefficients Feature Extraction and Backpropagation Neural Network," *International Conference on Science and Technology-Computer*, 2016

[11] S. Kathiravan *et al.*, "DAG-SVM based infant cry classification system using sequential forward floating feature selection," *Multidimensional Systems and Signal Processing*, pp. 961–976, 2017

[12] Y. Lavner, R. Cohen, D. Ruinskiy, H. Ijzerman, "Baby Cry Detection in Domestic Environment using Deep Learning," *I International Conference on the Science of Electrical Engineering*, pp. 1-5, 2016

[13] O. F. Reyes-Galaviz *et al.*, "Infant Cry Classification to Identify Hypoacoustics and Asphyxia with Neural Networks," *MICAI: Advances in Artificial Intelligence*, 2004

[14] X. Sevillano *et al.*, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds," *Applied Sciences*, 2016

[15] S. T. Aung *et al.*, "Analysis of Speech Features Extraction using MFCCs and PLP Extraction Schemes," *International Journal of Scientific Engineering and Technology Research*, vol. 06, pp. 2969–2973, 2017

[16] L. Muda *et al.*, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Technique," *Journal of Computing*, vol. 2, pp. 138–143, 2019

[17] D. Griffin, J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 236–243, 1984

[18] P. Vincent *et al.*, "Extracting and composing robust features with denoising autoencoders," *ICML*, 2008

[19] P. Vincent *et al.*, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *Journal of Machine Learning Research*, pp. 3371–3408, 2010

[20] K. He *et al.*, "Deep Residual Learning for Image Recognition," *CVPR,* 2015

[21] G. G. A. Celano, "A ResNet-50-based Convolutional Neural Network Model for Language ID Identification from Speech Recordings," *SIGTYP*, 2021

[22] S. Oprea *et al.*, " A Review on Deep Learning Techniques for Video Prediction" *arXiv*, 2020

[23] S. Zhang, M. Liu, J. Yan, "The Diversified Ensemble Neural Network," *NeurIPS*, 2020

[24] A. Krogh, J. Vedelsby, "Neural network ensembles, crossvalidation, and active learning," Ad*vances in Neural Information Processing systems*, pp. 231–238, 1995

BIOGRAPHIES OF AUTHORS

Joohyun Cha is an undergraduate student at at the University of Illinois Urbana-Champaign Grainger College of Engineering, where he is pursuing a degree in Computer Science (BS). His research interests include signal processing, computer vision, natural language processing, and pattern recognition. He can be contacted at email: joohyun6@illinois.edu

Gimin Bae received his M.S. degree in Department of Computer Engineering from Hanyang University, Seoul in 2019. Now, he is a Ph.D. student in Hanyang University. His research area includes Artificial Intelligence, Machine Learning, Anomaly Detection.