IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 23, 2020, accepted July 6, 2020, date of publication July 8, 2020, date of current version July 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007917

Traffic Density Classification Using Sound Datasets: An Empirical Study on Traffic Flow at Asymmetric Roads

KHAC-HOAI NAM BUI, HYEONJEONG OH, AND HONGSUK YI
Korea Institute of Science and Technology Information, Daejeon 34141, South Korea

Corresponding author: Hongsuk Yi (hsyi@kisti.re.kr)

This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea Ministry of Science and ICT (MSIT) (No.2018-0-00494, Development of deep learning-based urban traffic congestion prediction and signal control solution system) and Korea Institute of Science and Technology Information (KISTI) grant funded by the Korea Ministry of Science and ICT (MSIT) K-20-L02-C09-S01).

ABSTRACT Recently, with the rapid growth of Deep Learning models for solving complicated classification problems, urban sound classification techniques have been attracted more attention. In this paper, we take an investigation on how to apply this approach for the transportation domain. Specifically, traffic density classification based on the road sound datasets, which have been recorded and preprocessed on the urban road network, is taken into account. In particular, state-of-the-art methods for analyzing and extracting sound datasets have taken into account for the classification problem of traffic flow. Consequently, this study focuses on three main processes which are: i) generating image representation for the sequences of the road sound datasets; ii) proposing a convolutional neural network model for the feature extraction; iii) adopting a hybrid approach for the classification stage by combining convolutional neural network with other machine learning models. Regarding the experiment, the road sound dataset has been collected at an urban asymmetric road with different time periods (e.g., morning and evening) in order to evaluate our proposed method. Specifically, the implementations show promising results in which the accuracies are able to achieve from 92% to 95% for classifying traffic densities with different time periods.

INDEX TERMS Intelligent transportation system, traffic density classification, urban sound classification, deep learning, convolutional neural network.

I. INTRODUCTION

Traffic flow analysis is regarded as an important step for the development of the Intelligent Transportation System (ITS). Specifically, understanding the traffic patterns (e.g., volume, speed, and density) enables the traffic managers to provide smart services such as dynamic traffic light control [1], path planning [2], and abnormal detection [3]. Recently, Deep Learning (DL) models have achieved great success to overcome the complex problems of road traffic datasets [4]. Specifically, well-known DL models such as Deep Neural Network (DNN) [5], Recurrent Neural Network (RNN) [6], Convolutional Neural Network (CNN) [7], and Deep Reinforcement Learning (DQN) [8], [9] have been adopted for various applications in ITS as shown in Fig. 1.

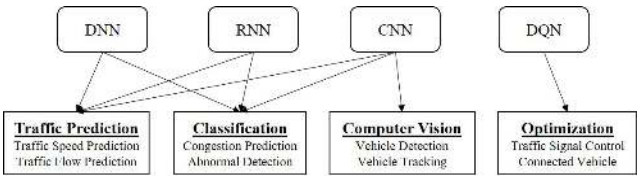

FIGURE 1. The applications of DL models in ITS.

Accordingly, regarding image classification problems, CNN is the most popular neural network model since the capability of this model for feature extraction and classification parts. Fig. 2 depicts an example of a CNN architecture for MNIST datasets. Specifically, the CNN-based methods follow a hierarchical architecture to build a trained model. Subsequently, the output result is defined based on the fully-connected layer. Consequently, CNN-based image classification has been applied in various domains.

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Chen.

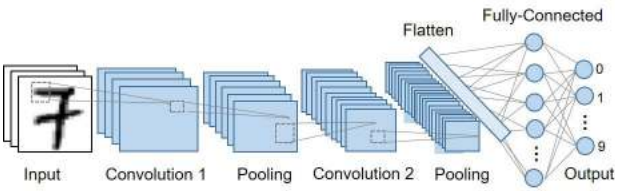

FIGURE 2. A CNN architecture to classify handwritten digits.

Reference [10] proposed an improved algorithm for CNN to assist medical experts for the breast cancer diagnosis problem. Reference [11] designed a 13-layer CNN for classifying the image-based fruit category. Reference [12] presented a CNN-base architecture for environmental sound classification problems. In the case of the transportation domain, [13] proposed a new method for large-scale traffic analysis by developing CNN architecture for traffic image datasets which are covered from network traffic.

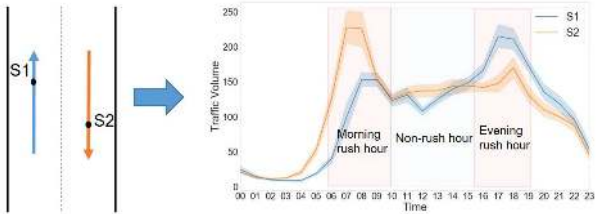

FIGURE 3. An example of measuring traffic volume at an asymmetric road using VDS.

In this paper, we focus on developing a CNN-based method for classifying the traffic density problem in which the input data are recorded and pre-processed from the traffic sound on the urban roads. Specifically, traffic conditions are conventionally represented by three patterns such as volume, speed, and density. Technically, surveillance systems (e.g., loop detectors) are set to measure traffic patterns. Fig. 3 shows the traffic volume of an asymmetric road using Vehicle Detection System (VDS) [14]. However, measuring traffic condition from surveillance systems have several drawbacks as follows:

- Surveillance systems are able to measure traffic volume and speed. However, measuring the traffic density is difficult since it depends on the spatial-temporal correlation [15].
- The time intervals depend on the sampling resolution of the detection devices, which are usually around a few minutes [13]. However, in several specific cases (e.g., traffic flow at the complex intersection), narrow intervals (e.g., few seconds) are required to classify the density of traffic conditions.
- Classifying detailed traffic patterns (e.g., traffic density of two directions in the same road) is difficult, which requires many involved sensors to collect the traffic data [16].

Particularly, with the rapid growth of DL models, analyzing traffic flow from low-cost video surveillance (CCTV) systems becomes a promising solution [17]. However, this approach is still very challenging in terms of improving accuracies of vehicle detection and tracking processes and computation time [18]. Therefore, in this paper, the urban sound classification approach is taken into account for the traffic density problem. Technically, urban sound classification is an emergent research topic that focuses on classifying environmental sounds form different objects (e.g., Air Conditioner, Car Horn, Children Playing, and so on). However, classifying the same objects with different time intervals (i.e., road sound of traffic flow) requires a specific structure of the feature extraction task. Hence, CNN-based methods have been adopted as one of the most promising models for extracting features in terms of time and frequency representations [19]. Regarding the sound datasets, we have collected the traffic sound data at certain points of the road network (e.g., the same location with the road sensors for verifying the results of the proposed approach) and preprocessed the sound excerpts with narrow intervals (i.e., 4 seconds). Therefore, by analyzing the collected traffic sound dataset, we are able to fix the enumerated drawbacks of surveillance systems for measuring traffic conditions. Specifically, the main contribution of this study is threefold as follows:

- We present a traffic density classification problem based on road sound datasets, entitled Road Sound Density Classification (RSDC). To the best of our knowledge, this is the first study by applying urban sound classification approach for traffic density classification problem. Specifically, the proposed approach is able to provide traffic conditions with narrow time intervals, which enables smart applications for ITS (e.g., dynamic traffic light control).
- We propose a new CNN architecture for the feature extraction process in which the model structure is not too deep and the results are comparable with well-known pre-trained CNN models (e.g., AlexNet and VGGNet). Furthermore, deep feature classification using well-known Machine Learning (ML) models such as K-Nearest Neighbor(KNN), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB) are taken into account to improve the classification performance.
- Regarding the experiment, a road sound dataset from an asymmetric road at a certain urban area is considered to evaluate the proposed approach. Particularly, we have collected and pre-processed around 14,000 labeled sound excerpts, entitled *RoadSound14k* dataset, which is classified into 6 classes of traffic densities. Specifically, we defined the traffic densities of urban roads into three periods of a day which are *Morning rush hour*, *Non-rush hour*, and *Evening rush hours* (as shown in Fig. 2). Consequently, the source code and dataset of this study are published and available to access.[1]

[1]https://github.com/BuiKhacHoaiNam/RoadSoundDensityClassification-RSDC-

The rest of this paper is structured as follows: Section 2 presents the background of urban sound classification techniques. The methodology of the RSDC problem is proposed in Section 3. The experiment is presented in Section 4 in which we collected and pre-processed the road sound datasets at a certain urban area to evaluate our proposed method. Section 5 includes the conclusion and future work of this study.

II. BACKGROUND

Urban sound classification is an emergent research topic with numerous real-world applications. Specifically, recent advancements in the field of image classification using DL models (e.g., CNN) enable the capability to classify the urban sound with high accuracy. Fig. 4 illustrates the main processes for urban sound classification.

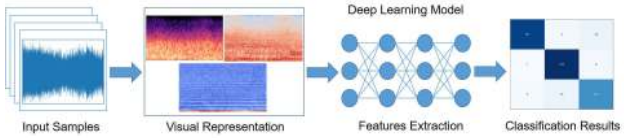

FIGURE 4. The main process of the urban sound classification problem.

Particularly, audio samples are processed in the readable format (e.g.,.wav) with a few seconds of the time duration (e.g., 4 seconds). Each sample is the amplitude of the wave (waveform) which is regarded as an input audio sample. Subsequently, the input data are converted into time-frequency images, which are pre-processed as the input of DL models for the classification. Specifically, there are two well-known methods for the visual representation process which are Spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) [20]. Technically, the main difference between the two methods is that the spectrogram adopts a linear spaced frequency scale (i.e., Short Time Fourier Transform (STFT)) and MFCC uses a quasi-logarithmic spaced frequency scale. For instance, Fig. 5 depicts an output image of the visual representations process from an input sound wave of the two methods.

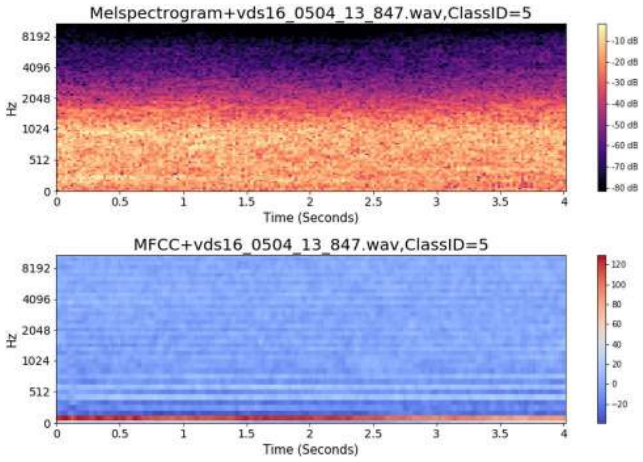

FIGURE 5. The output images by using Mel-Spectrogram and MFCC.

Regarding the urban sound datasets, *UrbanSound8K* [21] is a well-known open dataset that has been used for many studies in this research field [22]–[24]. Specifically, the dataset contains 8732 labeled sound excerpts in WAV format from 10 different classes. Consequently, in this study, our *RoadSound14k* dataset is preprocessed (e.g., audio files and meta-data file) following the format of *UrbanSound8K* dataset.

For the feature extraction and classification process, CNN architectures have proved the capability for the sound classification problem [25]. Specifically, several popular CNN architectures such as AlexNet, VGG, Inception, and ResNet have shown promising performances for the sound classification applications [26]. However, the most concern of adopting those aforementioned models is the computational cost. Specifically, the deep network structures of pre-trained CNN models are proposed to deal with the large size of input images (e.g., 224×224), which is not necessary in some cases of the sound classification problem. In the case of classification stage, recent studies focus on a hybrid approach that combines pre-trained CNN models for feature extraction and the conventional algorithms (e.g., SVM and KNN) for improving the performance of classification [27], [28].

III. ROAD SOUND DATASET-BASED TRAFFIC DENSITY CLASSIFICATION

This study proposes a traffic density classification approach using road sound datasets. Specifically, with inputs are audio samples of a few seconds of time duration, the objective is to determine which traffic condition they belong to with a corresponding classification accuracy score.

A. PROBLEM DESCRIPTION

Supporting $\mathcal{C} = \{c_1, c_2, \ldots c_n\}$ denotes a set of traffic condition labels a certain road, which is defined based on the time intervals of a day (e.g., Morning rush hour, Non-rush hour, and Evening rush hour). The traffic density classification problem is regarded as a supervised learning problem in which given a training set $\mathcal{X} = \{(x_i, Y_i \mid 1 \leq i \leq m)\}$, where $Y_i \subset \mathcal{C}$, the objective is to learn a multi-label classifier from $\mathcal{X}$ to predict labels of new audio samples. Specifically, the classification problem using CNN model with parameter Θ can be formulated as follows:

$$\mathcal{F}(\mathcal{X} \mid \Theta) = f_n(\ldots f_2(f_1(\mathcal{X} \mid \theta_1) \mid \theta_2) \mid \theta_n) \quad (1)$$

where $f_j(\mathcal{X} \mid \theta_i)$ (j $\in$ n) represents the layer j^{th} of the network with total number of layer n.

B. METHODOLOGY

Fig. 6 demonstrates the pipeline of the proposed method for the RSDC problem. Specifically, the method includes three main processes which are image representation using time-frequency spectrogram, deep feature extraction using a new CNN architecture, and classification stage using well-known ML algorithms. More detail of the processes are sequentially described as follows:

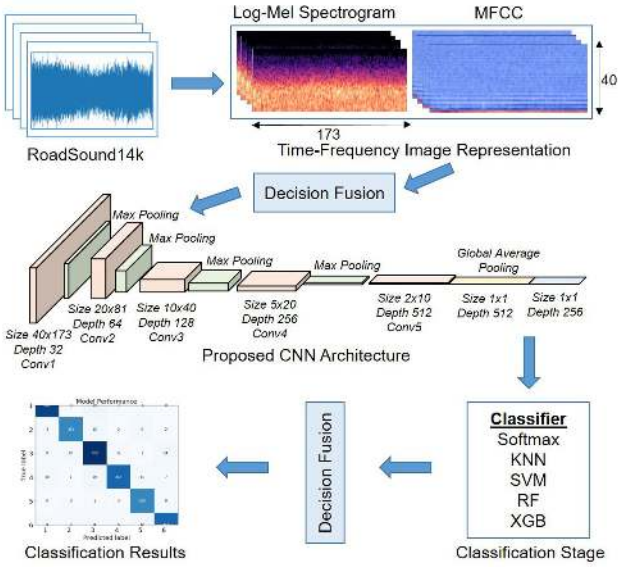

FIGURE 6. System architecture for the RSDC problem.

1) IMAGE REPRESENTATION

As mentioned above, log-mel spectrogram and MFCC are popular methods which have widely adopted for sound recognition. Therefore, in this study, two aforementioned methods are taken into account for the time-frequency representations [20]. Technically, the methods can be executed by using Librosa library [29]. Specifically, the main steps of this process are expressed as follows:

- Determining the window size (n_fft = 2048) and hop length (hop_length = 512).
- Computing STFT to transform from time domain to frequency domain which is formulated as follows:

$$\mathcal{X}(\tau, \omega) = \int_{-\infty}^{\infty} x(i)\omega(i-\tau)e^{-j\omega i}di \tag{2}$$

  where $x(i)$ and $\omega(\tau)$ represent the input sample and Hanning window, respectively.
- Generating Mel-scale spectrogram by determining Mel-scale value (n_mels = 40), converting sound intensity to log amplitude (mel_db) and normalizing the mel_db values (normalized_mel $\in [-1,1]$). In the case of MFCC, Discrete Cosine Transform (DCT) is adopted to the logarithm of Mel-spectrogram features and generated the compressed representation of Mel-frequencies.

Consequently, the dataset is a set of spectrogram images in which each image represents an audio sample. For instance, Fig. 7 illustrates the sequential results of this process from an input sample.

2) CNN MODEL

The next process is to develop a CNN model with the generated dataset for making the predictions. In this study, instead of adopting well-known CNN architecture such as VGG or ResNet for training the road sound dataset, we develop and implement our own CNN model for the feature extraction. Fig. 8 shows the model structure of our proposed method. Accordingly, the proposed architecture comprises 5 convolutional layers interleaving with 4 max-pooling operations and 1 fully connected layers. Specifically, a convolutional layer can be expressed as follows:

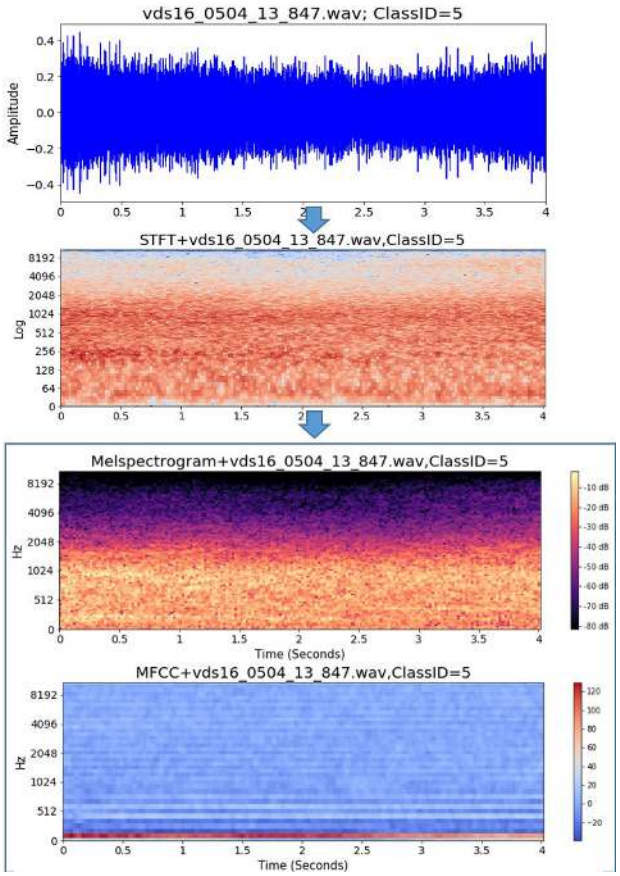

FIGURE 7. The results of visual representation process using log-Mel spectrogram and MFCC methods.

$$f_i(\mathcal{X}_i \mid \theta_i) = h(W * \mathcal{X}_i + b) \tag{3}$$

where W represents the collection of M 3-D kernels (filters). b and h(.) refer to the bias term and activation function, respectively. Regarding the activation function, redirected linear unit (ReLU) is adopted with the convolutional layers, which is formulated as follows:

$$h(x) = max(0, x) \tag{4}$$

Furthermore, instead of using the max-pooling in the last layer and flattened, we adopt the Global Average Pooling (GAP) to deal with the overfitting problem by reducing the number of parameters [30]. Therefore, only two fully connected layer (Dense) is applied in our proposed architecture. Specifically, the detailed architecture is described as follows:

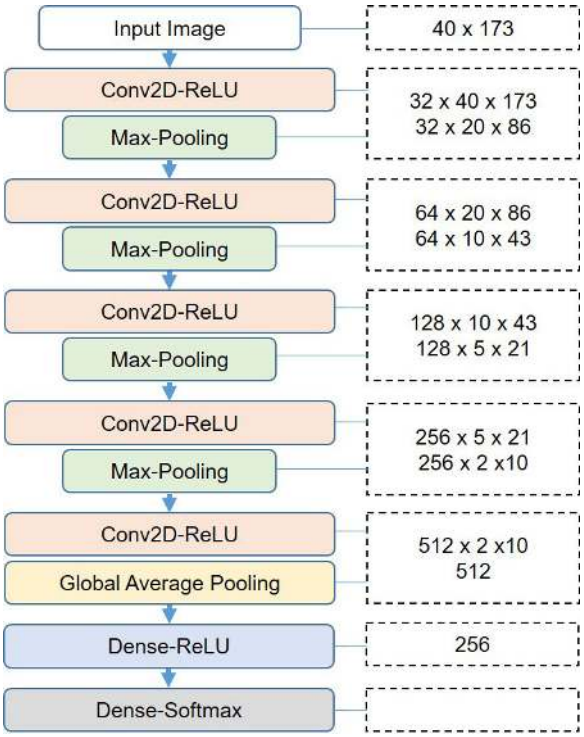

FIGURE 8. Proposed CNN architecture with Softmax for the classification.

- Layer 1: Filters=32; Kernel=(3,3); Activation = ReLU; Max pooling = (2,2).
- Layer 2: Filters=64; Kernel=(3,3); Activation = ReLU; Max pooling = (2,2).
- Layer 3: Filters=128; Kernel=(3,3); Activation = ReLU; Max pooling = (2,2).
- Layer 4: Filters=256; Kernel=(3,3); Activation = ReLU; Max pooling = (2,2).
- Layer 5: Filters=512; Kernel=(3,3); Activation = ReLU.

3) CLASSIFICATION STAGE

Softmax function with the cross-entropy (loss function) is generally utilized in CNNs for the multiclass classification with a posterior probability output in which the function can be computed as follows:

$$f_i(z) = \frac{e^{z_i}}{\sum_j^{\mathcal{C}} e^{z_j}} \tag{5}$$

where z_i represent the scores that are inferred by the net of each class in $\mathcal{C}$. Therefore, the loss function uses the form of cross-entropy (CE) loss is formulated as follows:

$$Loss_{CE} = -log(\frac{e^{z_p}}{\sum_j^{\mathcal{C}} e^{z_j}}) \tag{6}$$

where z_p represents the model score for the positive class. Recent studies take an investigation on a hybrid approach that combines CNN for feature extraction and other ML models instead of using softmax function for the classification process [31]–[33]. Consequently, in this study, we take into account several standard ML models in literature for the classification process and using Softmax as the based line method for the classification. Specifically, the algorithms for the classification are described as follows:

- KNN: is regarded as one of the simplest classifiers in which the algorithm requires two main parameters (i.e., distance metric and the value of neighbors) [28].
- SVM: is a kind of linear classifier that is able to extract data in the form of N-dimensional vectors. Specifically, the objective of this algorithm is to find the optimal solution for the simple linear mapping [34].
- RF: is an ensemble learning algorithm using tree-type classifiers in which the number of trees is the most important parameter [35].
- XGBoots: is another decision tree ensembles method. Specifically, in this algorithm, individual trees are generated by using multiple cores, and data is organized for minimizing the lookup time in order to reduce the training time and improve the accuracy of the classification [36].

IV. EXPERIMENT

A. DATA DESCRIPTION

For the experiment, *RoadSound14k* dataset is collected on the main road of an urban area. Particularly, in order to learn more detail of the traffic pattern, the data is recorded and pre-processed at an asymmetric road as shown in Fig. 3. Furthermore, traffic conditions in each direction are determined into three classes which are Morning rush hour, Non-rush hour, and Evening rush hour. Consequently, there are 6 classes of the traffic condition in this study which are explained in more detail in the Tab. 1.

TABLE 1. ***RoadSound14k*** **dataset.**

Condition	Direction	ClassID	Time Interval	Samples
Morning rush hours	A	1	7:00 - 10:00	2646
Non-rush hours	A	2	13:00 - 15:00	1800
Evening rush hours	A	3	17:00 - 20:00	2700
Morning rush hours	B	4	7:00 - 10:00	2618
Non-rush hours	B	5	13:00 - 15:00	1791
Evening rush hours	B	6	17:00 - 20:00	2700
Total Samples				14255

Specifically, the *RoadSound14k* dataset contains 14255 audio samples in which the time duration in each file is around 4 seconds. The audio files are recorded with 48 kHz and the number of samples is different in each class. For the training model, we divide the dataset with 70% data to train (training data), 10% of the validation set, and 20% for testing data.

B. EXPERIMENTAL SETUP

Tab. 2 illustrates the parameter that we use for the experiment. Specifically, for the image representation process, the value of window size and hop size are 2047 and 512, respectively. The size of input images for the CNN model is 40×173.

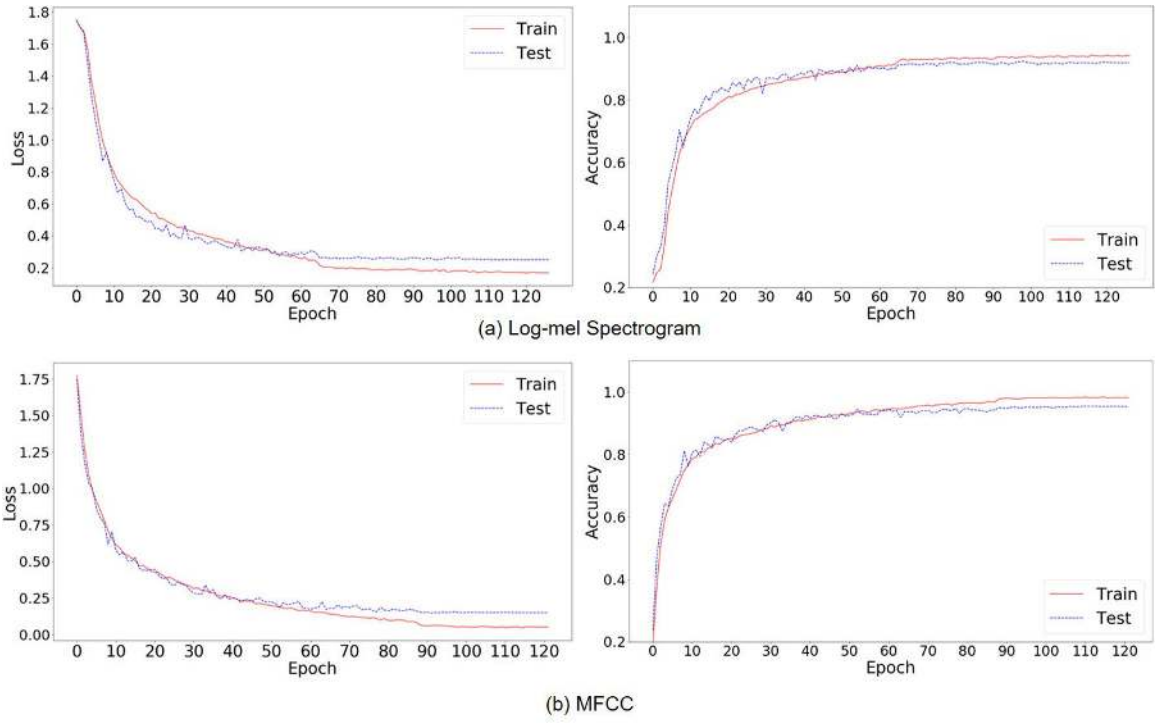

FIGURE 9. The comparison results between two image representation methods.

TABLE 2. Experiment parameters.

Parameter	Value
Window Size/Type	2048/Hamming
Hop Length	512 samples
Input Image Size	40x173
Initial Learning Rate	$1e^{-3}$
Optimizer	Adam
Batch Size	64
Epoch	200

Regarding the training process for feature extraction using the proposed CNN model, we adopt Adam optimization [37] with the number of the batch size is 64. The initial learning rate starts with $1e^{-3}$ (minimum $1e^{-6}$) with 200 epochs. Dropout (Default = 0.5) was applied during the training process in each layer to reduce the overfitting [38]. Furthermore, Early stopping is performed by monitoring the validation error. The network is executed in Python with Tensorflow as the back-end [39] and works well by a PC with Core i7 16-GB CPU and 32GB GPU memories in which we used the GPU for acceleration.

For the evaluation, predicted results are compared with the testing data to evaluate the performance metrics. Furthermore, the evaluation criteria include accuracy, precision, recall, and F1-score, by using the confusion matrix, which is sequentially calculated as follows:

$$Accurancy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Prescision + Recall} \tag{10}$$

where TP, TN, FP, FN refer to the terms used in the confusion matrix which are True Positive, True Negative, False Positive and False Negative, respectively.

C. RESULT ANALYSIS

1) THE RESULT OF TIME-FREQUENCY REPRESENTATION

In order to evaluate the effectiveness of the proposed approach, three implementations are taken into accounts such as the effectiveness of the visual representation, the CNN architecture for feature extraction, and the classification stage. Fig. 9 shows the results with the proposed CNN using softmax for the classification of the two image representation methods. As shown in the figures, using MFCC outperforms the log-Mel-spectrogram method. Specifically, the log-Mel spectrogram features involve highly correlation which is the cause of reducing the performance in some specific models. Moreover, in the case of MFCC, we remain the Mel-coefficients without dropping ($n_{mfcc} = 40$) since this study using neural networks for feature extraction. Therefore, for the rest of the evaluation, we utilize the MFCC method for the image representation process to deal with the RSDC problem.

2) THE RESULT OF CNN ARCHITECTURE

In order to evaluate the effectiveness of the proposed CNN model for the feature extraction process. We compare the

TABLE 3. Comparison results among CNN architectures.

Evaluation Critical	AlexNet	VGG-16	VGG-19	Proposed CNN
Accuracy	0.9273	0.9277	0.9281	0.9372
Precision	0.9288	0.9284	0.9284	0.9378
Recall	0.9274	0.9277	0.9281	0.9372
F1-Score	0.9276	0.9278	0.9282	0.9373
Time(Sec)	398	830	1022	274

proposed architecture with well-known CNN architecture such as AlexNet, VGG-16, and VGG-19 in which the softmax is used for the classification. Tab. 3 shows the results with different evaluation critical. Accordingly, our proposed CNN architecture is able to achieve a better result on the *RoadSound14k* dataset with the least time-consuming of the training process. Specifically, Fig. 10 shows the classification result with Softmax algorithm for *RoadSound14k* dataset using our approach.

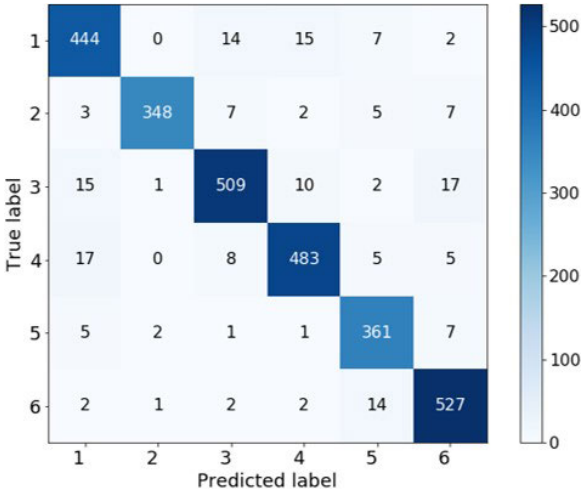

FIGURE 10. The prediction results using the proposed CNN model with softmax for the classification.

3) THE RESULT OF CLASSIFICATION ALGORITHMS

As we mentioned above, in the classification stage, several well-known ML models are taken into account to improve the performance of accuracy. Consequently, Fig. 11 illustrates the performance by adopting other ML models for the classification instead of using the Softmax algorithm. In this regard, the proposed CNN architecture is utilized as

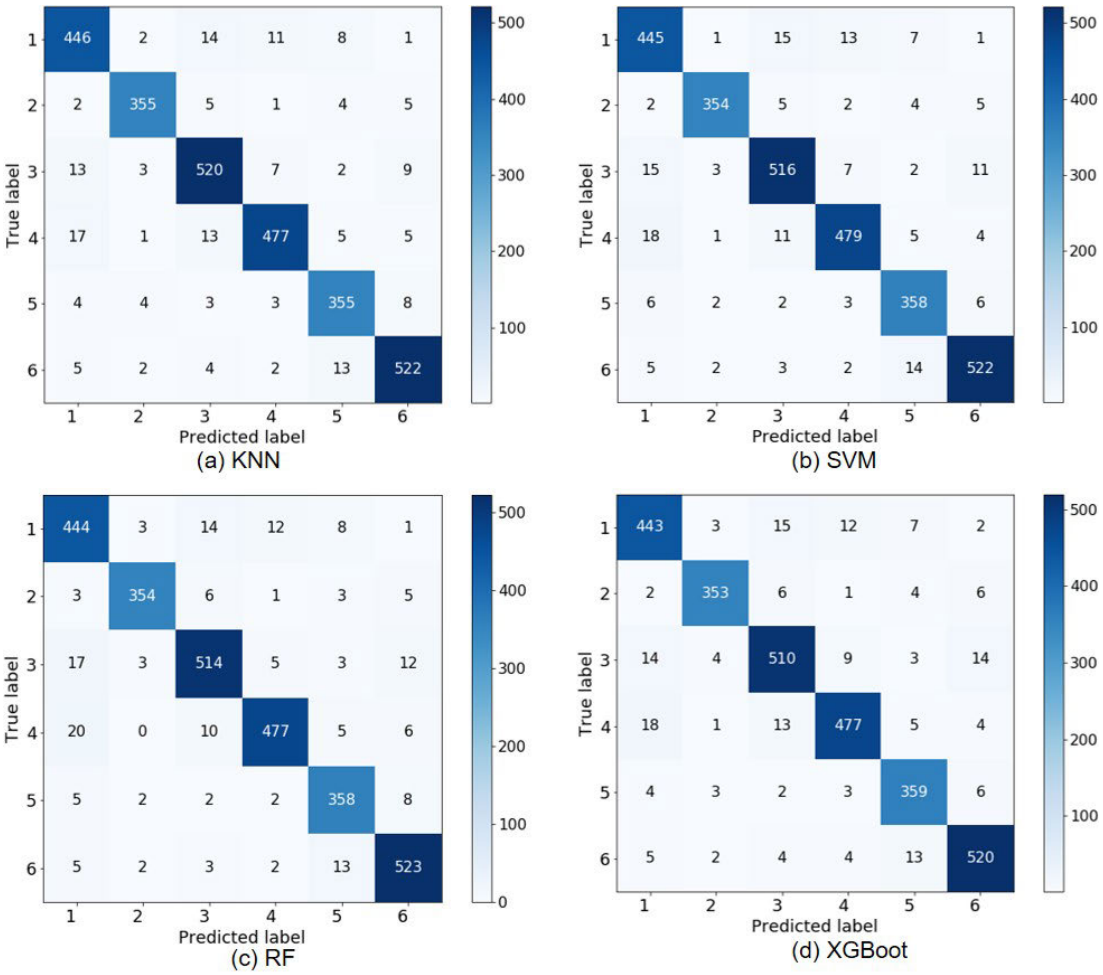

FIGURE 11. The classification results using different ML models.

the pre-trained model for feature extraction. As results, using the hybrid CNN-KNN model is slightly better than other methods such as CNN, CNN-SVM, CNN-RF, and CNN-XGboot in terms of training the *RoadSound14k* dataset. However, the drawback of KNN is time-consuming in which the algorithm costs around 40 times for training the features comparing with the XGboot algorithm. Specifically, KNN is robust and effective in terms of training the noisy and large number of input data. Nevertheless, we need to compute the distance of all training samples which is the cause of high computation time. More detail of this issue is shown in Tab. 4.

TABLE 4. Accuracy results (%) of different algorithms for the classification stage.

ClassID	Softmax	KNN	SVM	RF	XGBoot
1	92.12	**92.53**	92.32	92.12	91.91
2	93.55	**95.43**	95.16	95.16	94.89
3	91.88	**93.86**	93.14	92.78	92.06
4	**93.24**	92.08	92.47	92.08	92.08
5	**95.75**	94.16	94.96	94.96	95.22
6	**96.17**	95.25	95.25	95.44	94.89
Average	93.72	**93.83**	93.79	93.65	93.37
Time(Sec)	-	638	45	333	**16**

Consequently, the detailed result in each class of the proposed approach for the RSDC problem is shown in Tab. 5. Specifically, the MFCC method is adopted for the image presentation process. Then, the proposed CNN architecture with 5 convolutional layers is executed as the pre-trained model for the feature extraction. Sequentially, the KNN algorithm is applied in the classification stage to improve the performance of the classification process.

TABLE 5. Classification results for the RSDC problem of the proposed CNN-KNN model.

ClassID	Precision	Recall	F1-Score	Support
1	0.92	0.93	0.92	482
2	0.97	0.95	0.96	372
3	0.93	0.94	0.93	554
4	0.95	0.92	0.94	518
5	0.92	0.94	0.93	377
6	0.95	0.95	0.95	548
micro avg	0.94	0.94	0.94	2851
macro avg	0.63	0.63	0.63	2851
weight avg	0.94	0.94	0.94	2851

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a new approach for traffic density classification using the road sound datasets, entitled RSDC problem, which is inspired by the recent advances of urban sound classification problem using CNN-based methods. Specifically, traffic sound data is collected on the main roads with different time intervals for the classification of the traffic conditions. Particularly, a new CNN architecture including 5 convolutional layers is proposed for the feature extraction process. Then, several well-known ML models are implemented in order to improve the performance of accuracy. The experiment indicates the promising results of our method for traffic condition classification using road sound datasets.

From our point of view, there are two issues that we are taking into account regarding the future work of this study: i) increasing the size of the *RoadSound14k* dataset which involves various traffic patterns (e.g., daily, weekend, and weather conditions); ii) applying the proposed method for smart applications of transportation (e.g, dynamic traffic light control). Specifically, analyzing traffic density using road sound datasets is able to provide the traffic condition in a short time (few seconds) which can be applied for the dynamic traffic light control to improve the traffic flow in complex areas.

REFERENCES

[1] K.-H.-N. Bui, J. E. Jung, and D. Camacho, "Game theoretic approach on real-time decision making for IoT-based traffic light control," *Concurrency Comput., Pract. Exper.*, vol. 29, no. 11, p. e4077, Jun. 2017.
[2] K.-H.-N. Bui and J. J. Jung, "ACO-based dynamic decision making for connected vehicles in IoT system," *IEEE Trans. Ind. Informat.*, vol. 15, no. 10, pp. 5648–5655, Oct. 2019.
[3] K.-H.-N. Bui, S. Cho, J. J. Jung, J. Kim, O.-J. Lee, and W. Na, "A novel network virtualization based on data analytics in connected environment," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 1, pp. 75–86, Jan. 2020.
[4] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 24, 2019, doi: 10.1109/TITS.2019.2929020.
[5] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.
[6] H. Yi and K.-H.-N. Bui, "An automated hyperparameter search-based deep learning model for highway traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 24, 2020, doi: 10.1109/TITS.2020.2987614.
[7] T. Pamula, "Road traffic conditions classification based on multilevel filtering of image content using convolutional neural networks," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 3, pp. 11–21, Jun. 2018.
[8] Z. Ning, Y. Li, P. Dong, X. Wang, M. S. Obaidat, X. Hu, L. Guo, Y. Guo, J. Huang, and B. Hu, "When deep reinforcement learning meets 5G-enabled vehicular networks: A distributed offloading framework for traffic big data," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1352–1361, Feb. 2020.
[9] Z. Ning, K. Zhang, X. Wang, S. Mohammad Obaidat, L. Guo, X. Hu, B. Hu, Y. Guo, B. Sadoun, and R. Y. Kwok, "Joint computing and caching in 5g-envisioned Internet of vehicles: A deep reinforcement learning-based traffic control system," *IEEE Trans. Intell. Transp. Syst.*, early access, Feb. 5, 2020, doi: 10.1109/TITS.2020.2970276.
[10] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification," *Expert Syst. Appl.*, vol. 120, pp. 103–115, Apr. 2019.
[11] Y.-D. Zhang, Z. Dong, X. Chen, W. Jia, S. Du, K. Muhammad, and S.-H. Wang, "Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3613–3632, Feb. 2019.
[12] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, Apr. 2019.
[13] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, Apr. 2017.
[14] H. Yi and K.-H. N. Bui, "VDS data-based deep learning approach for traffic forecasting using LSTM network," in *Proc. 19th EPIA Conf. Artif. Intell. (EPIA)*, Sep. 2019, pp. 547–558.
[15] J. Chung and K. Sohn, "Image-based learning to measure traffic density using a deep convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1670–1675, May 2018.
[16] C. Meng, X. Yi, L. Su, J. Gao, and Y. Zheng, "City-wide traffic volume inference with loop detector data and taxi trajectories," in *Proc. 25th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2017, pp. 1:1–1:10.

[17] K.-H.-N. Bui, H. Yi, and J. Cho, "A multi-class multi-movement vehicle counting framework for traffic analysis in complex areas using CCTV systems," *Energies*, vol. 13, no. 8, p. 2036, Apr. 2020.
[18] M. Naphade, S. Wang, C. D. Anastasiu, Z. Tang, M. Chang, X. Yang, L. Zheng, A. Sharma, R. Chellappa, and P. Chakraborty, "The 4th AI city challenge," *CoRR*, vol. abs/2004.14619, pp. 1–10, Apr. 2020.
[19] H. Yoonchang, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, Nov. 2017, pp. 1–5.
[20] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *CoRR*, vol. abs/1706.07156, pp. 1–5, Jun. 2017.
[21] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia (MM)*, Nov. 2014, pp. 1041–1044.
[22] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2017, pp. 1–5.
[23] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," in *Proc. 21st Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst. (KES)*, Sep. 2017, pp. 2048–2056.
[24] B. McMahan and D. Rao, "Listening to the world improves speech command recognition," in *Proc. 32nd AAAI Int. Conf. Artif. Intell. (AAAI)*, Feb. 2018, pp. 378–385.
[25] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.
[26] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.
[27] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[28] F. Demir, D. A. Abdullah, and A. Sengür, "A new deep CNN model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66529–66537, 2020.
[29] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf. (SciPy)*, 2015, pp. 18–24.
[30] S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing*, vol. 219, pp. 88–98, Jan. 2017.
[31] A. F. Agarap, "An architecture combining convolutional neural network (CNN) and support vector machine (SVM) for image classification," *CoRR*, vol. abs/1712.03541, pp. 1–4, Feb. 2017.
[32] F. Demir, V. Bajaj, M. C. Ince, S. Taran, and A. engür, "Surface EMG signals and deep transfer learning-based physical action classification," *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8455–8462, Dec. 2019.
[33] U. Knauer, C. S. von Rekowski, M. Stecklina, T. Krokotsch, T. Pham Minh, V. Hauffe, D. Kilias, I. Ehrhardt, H. Sagischewski, S. Chmara, and U. Seiffert, "Tree species classification based on hybrid ensembles of a convolutional neural network (CNN) and random forest classifiers," *Remote Sens.*, vol. 11, no. 23, p. 2788, Nov. 2019.
[34] Y. Tang, "Deep learning using support vector machines," *CoRR*, vol. abs/1306.0239, pp. 1–6, Feb. 2013.
[35] T. Li, J. Leng, L. Kong, S. Guo, G. Bai, and K. Wang, "DCNR: Deep cube CNN with random forest for hyperspectral image classification," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3411–3433, Feb. 2019.
[36] H. T. Weldegebriel, H. Liu, A. U. Haq, E. Bugingo, and D. Zhang, "A new hybrid convolutional neural network and eXtreme gradient boosting classifier for recognizing handwritten Ethiopian characters," *IEEE Access*, vol. 8, pp. 17804–17818, 2020.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[39] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.0446, pp. 1–19, Mar. 2016.

KHAC-HOAI NAM BUI received the M.S. degree in computer science and information engineering from Aletheia University, New Taipei City, Taiwan, in 2014, and the Ph.D. degree in computer science and engineering from Chung-Ang University, Seoul, South Korea, in 2018.

He has been a Researcher with the Korea Institute of Science and Technology Information (KISTI), Daejeon, South Korea, since February 2019. Before joining KISTI, he was a Postdoctoral Researcher with Chung-Ang University, in September 2018. His research interests include the Internet of Things and ambient intelligence by using AI methodologies such as data mining, machine learning, and logical reasoning. Recently, he focuses on applying deep learning models for smart applications, such as transportation and home energy management systems.

HYEONJEONG OH received the B.S. degree in computer science and engineering from Chungnam National University, South Korea, in February 2020. She has been doing an Internship with the Korea Institute of Science and Technology Information (KISTI), South Korea, since April 2020. Her research interest includes big data analysis for smart city applications.

HONGSUK YI received the Ph.D. degree from Sogang University, Seoul, South Korea, in 1997. He has been a Principal Researcher with the Korea Institute of Science and Technology Information (KISTI), South Korea, since 2000. His research interests include traffic congestion problem, smart cities, supercomputing, and heterogeneous computing by using deep learning techniques. Recently, his research topic focuses on developing traffic signal control solutions by using deep learning methodologies.