Safety Monitoring System of Personal Mobility Driving Using Deep Learning

Eunji Kim[1], Hanyoung Ryu[1], Hyunji Oh[1], and Namwoo Kang[2,*]

[1]*Sookmyung Women's University, Seoul, Korea*

[2]*Korea Advanced Institute of Science and Technology, Daejeon, Korea*

*corresponding author: nwkang@kaist.ac.kr

Abstract

Although the e-scooter sharing service market is growing as a representative last-mile mobility, the accident rate is increasing proportionally as the number of users increases. This study proposes a deep learning-based personal mobility driver monitoring system that detects inattentive driving by classifying vibration data transmitted to the e-scooter when the driver fails to concentrate on driving. First, the N-back task technique is used. The driver was stimulated by external visual and auditory factors to generate a cognitive load, and vibration data were collected through a six-axis sensor. Second, the generated vibration data were preprocessed using short-time Fourier transform (STFT) and wavelet transform (WT) and then converted into an image (spectrogram). Third, four multimodal convolutional neural networks (CNNs) such as LeNet-5, VGG16, ResNet50, and DenseNet121 were constructed and their performance was compared to find the best architecture. Experimental results show that multimodal DenseNet121 with WT can accurately classify safe, slightly anxious, and very anxious driving conditions. The proposed model can be applied to real-time monitoring and warning systems for sharing service providers and used as a basis for insurance and legal action in the case of accidents.

Keywords: Deep Learning, Personal Mobility, Short-time Fourier Transform (STFT), Wavelet Transform (WT), Convolutional Neural Networks (CNNs)

Note: A previous version of this manuscript was presented at the third Asia Pacific Conference of the Prognostics and Health Management Society (Jeju, Korea, Sep 8-11, 2021)

1. Introduction

Environmentally friendly means of transportation has attracted considerable attention due to the emergence of global warming and environmental problems. Accordingly, various electric-powered vehicles have been developed for transportation and the car-sharing service market that allows users to share vehicles without owning them has grown. In addition, for first-last-mile mobility, the personal mobility sharing service industry has developed portable, accessible, and eco-friendly vehicles, such as an electric bicycles or scooters (e-scooters) (Chong et al., 2011).

The personal mobility market is growing exponentially. The market research firm Berg Insight predicted that 774,000 units of shared e-scooters at the end of 2019 will increase to 4.6 million units by 2024 (Berg Insight, 2020). According to the report of Global Personal Mobility Devices Market, personal mobility will grow to a market value of $9.4 billion from 2016 to 2026. A total of 150,000 e-scooters have been operational in 177 cities in the United States and Europe since 2019 when the e-scooter sharing service emerged as a means of transportation for the first-last mile and its estimated market size is $740 million (Facts & Factors, 2020). According to data from the Korea Financial Supervisory Service, the number of operational units of 16,570 of 20 personal mobility sharing service companies since 2019 is expected to exceed 40,000 units by 2020 (Kwon, 2020). According to data from the Korea Transport Institute, the mobility market is expected to grow to 300,000 units by 2022 (The Korea Transport Institute, 2017).

As the number of personal mobility users increases, the accident rate is also increasing proportionally. According to Forbes magazine, from 2014 to 2018 in the U.S., about 3,300 patients were hospitalized for electric scooter-related injuries, a 365% increase. During the same period, total electric scooter-related injuries totaled 39,000, an increase of 222% (Mack, 2020). According to Stuff, New Zealand cost less than $15 million in taxes over two years due to electric scooter-related injuries. Between October 2018 and January 2021, a total of 6,284 incidents involving electric scooters were received, and paid $14.98 million. In January 2021, 200 accidents cost $458,703 (Hutt, 2021). According to the Korea Consumer Agency, the number of electric scooter accidents in Korea due to careless driving increased about 17 times from 14 cases in 2015 to 233 cases in 2018 (Korea Consumer Agency, 2019). And according to the insurance industry, 447 cases were received in 2019 (Kwon, 2020).

Each country amends its road traffic laws and reinforces safety measures to solve various causes, and e-scooter sharing service operators are also in the process of developing technologies for the safety of pedestrians and drivers. According to CNN News, the Swedish operator Voi has registered more than 6 million e-scooter riders in 50 European cities and collaborated with start-up company Luna to develop a deep learning system that can detect the road surface and nearby presence (Lewis, 2021). Spin, Ford's micromobility division, has recently announced that it will add computer vision and machine learning technology to its next-generation e-scooters (Lewis, 2021). Lime, a U.S. shared service provider, introduced a technology that uses speed and vibration patterns to identify

driving on sidewalks in 2020 (Lewis, 2021). Olulo, a Korean Kickgoing operator, developed a technology that recognizes pedestrians while driving using ultrasmall cameras in the front, back, and sides of e-scooters, applied automatic speed limits when entering sidewalks or protected areas and limited the boarding of two persons in the vehicle in 2020. The Sing Sing operator PUMP has developed a black box for e-scooters (Choi, 2020).

Although these technologies are protective measures for external impact, studies on driver carelessness are lacking. This study aims to analyze the driver's concentration level while driving and identify the section where cognitive loads occur. Human carelessness can be assessed in various ways. For example, anxiety can be measured through a person's brainwave or electrocardiogram using wearable sensors or a system can be utilized to locate the driver's visual direction to confirm whether the person is staring off the road while driving. However, wearable sensors are generally inconvenient and expensive, making them difficult to use for service.

This study hypothesizes that if a driver fails to concentrate on driving an e-scooter, then the e-scooter will vibrate in a different way than usual. Since vibration data can be easily collected from the driver's cellular phone or sensors mounted on the e-scooter, the monitoring system using the vibration data will help reduce the cost and achieve high effectiveness. The problem is how to identify the difference between the vibration data generated during safe driving and unsafe driving.

This research proposes a deep learning-based personal mobility driver monitoring system that detects inattentive driving by classifying vibration data transmitted to the e-scooter when the driver fails to concentrate on driving. First, a visual/auditory N-back task on drivers of e-scooters is conducted to collect vibration data from six-axis sensors due to cognitive load on the road. Second, short-time Fourier transform (STFT) and wavelet transform (WT) methods are used to convert vibration data into images. Third, multimodal convolutional neural networks (CNNs) such as LeNet-5, VGG16, ResNet50, and DenseNet121 are built to classify the safe, slight-anxiety, and high-anxiety driving states of drivers through preprocessed image data.

The remainder of this paper is organized as follows. Various studies related to this study and the methodology used are presented in Chapter 2. The overall framework used in this study and the resulting e-scooter experimental, data preprocessing, and deep learning methods are described in Chapter 3. The results of deep learning models are discussed in Chapter 4. Finally, conclusions and future research plans are provided in Chapter 5.

2. Related Works

2.1. Convolutional Neural Network

CNN is a deep learning model that maintains spatial information of images through convolutional and pooling layers and implements feature maps to find important features of images and perform

classification on a fully connected (FC) layer (Krizhevsky et al., 2012). CNNs have been widely used in prognostics and health management (PHM) research. For example, 2D CNN was used for gearbox failure signal detection by preprocessing time-frequency images (Wang et al., 2017). 1D CNN was also used for normalized vibration signals for real-time vibration-based damage detection and positioning without the need for image preprocessing (Abdeljaber et al., 2017).

In thus study, we use and compare advanced CNN models such as LeNet-5 (LeCun, 2015), VGG16 (Simonyan et al., 2014), ResNet50 (He et al., 2016), and DenseNet121 (Huang et al., 2017) to classify normal and abnormal vibration data. The LeNet-5 architecture has the simple architecture which consists of three convolutional layers, two average pooling layers, and one FC layer. VGG16 sets the filter size of the convolution kernel to 3×3 and investigated the effect of depth of the network. The results showed that the increase in depth of layers improves the performance. Therefore, these early CNN models have deepened the network and increased the layer to improve data accuracy. However, the deep architecture resulted in problems, such as vanishing/exploding gradients, and poor performance.

ResNet addresses these problems by adding the residual learning technique to an existing deep learning method as shown in Fig. 1. ResNet is a structure with $F(\mathbf{x}) + \mathbf{x}$ as the output when the input is $\mathbf{x}$. This process allows ResNet to construct 152 layers, and the ResNet model won the ImageNet Large-scale Visual Recognition Challenge for image classification, detection, and localization in 2015 as well as MS COCO Image Captioning Challenge for detection and segmentation in 2015 (He et al., 2016).

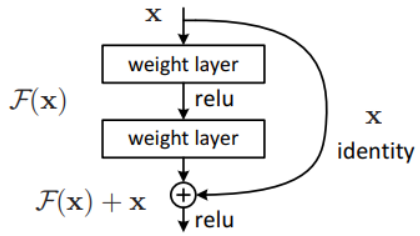

Figure 1: ResNet - residual learning (He et al., 2016)

While ResNet is a structure that sums up previous and current results, DenseNet is a structure that concatenates all results together (Fig. 2). The DenseNet model preserves and reflects all results of the entire layer so that it can reduce vanishing gradients. This model can also lower computational cost because it requires fewer parameters than conventional models (Huang et al., 2017).

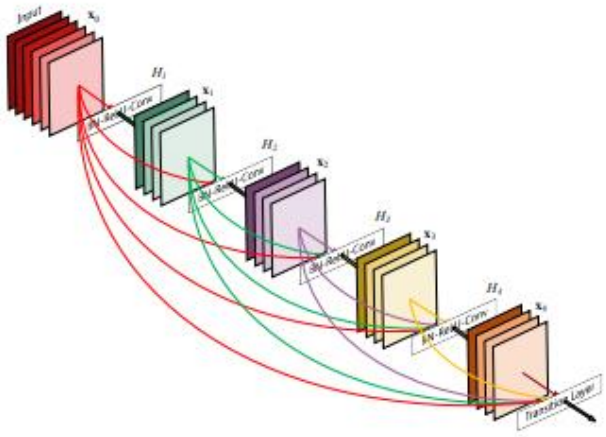

Figure 2: DenseNet - structure that concatenates all results (Huang et al., 2017)

2.2. Signal Preprocessing

STFT is a method that performs Fourier transform on data of local sections with respect to time and changes the original signal into a spectrogram to identify frequency bands and time changes visually. The classical Fourier transform is unsuitable for time-frequency domain analysis because it does not contain information in the local section. Meanwhile, STFT can determine the phase information of a signal that changes with time in the frequency or local range of a sine wave, divide a long signal into short units of the same length, and perform discrete Fourier transform (DFT) on each segment. STFT is useful for diagnosing faults in machines and signals or capturing the moment of a signal. Devadasu et al. (2016) used STFT to diagnose faults through voltage and current analysis. Khang et al. (2015) converted vibration signals recorded around them through STFT to diagnose defects or wear of mechanical parts. Thiruvaran et al. (2013) calculated derivatives of short-time magnitudes of different harmonics using STFT to capture the switching moment.

WT, a method used to express time and frequency components of a signal with changing components, removes or attenuates noise included in the original signal by dividing the original signal into minute signal waveforms and then calculating the signal strength of each waveform section. WT provides multiresolution analysis in time and frequency with high temporal resolution in the high-frequency domain and high frequency resolution in the low-frequency domain. WT is widely used to denoise data or extract error and feature values (Wei, C. et al., 2020).

Although STFT presents the advantage of easy interpretation, finding a specific signal in the presence of a considerable amount of noise can be challenging. In addition, time and frequency resolutions are also fixed given that STFT uses a fixed window size. Although STFT fails to improve both time and frequency resolutions simultaneously, WT complements this limitation by transforming the original signal into a scalogram to detect changes in the frequency band visually. Change timing is important for a rapidly changing signal, and the period or frequency of the change

is important for a slowly changing signal. WT is more effective than STFT in terms of time-frequency analysis because the temporal resolution changes with frequency (Jurado et al., 2002).

WT showed more stable classification performance with fewer iterations than STFT in a study using WT and CNN to diagnose gearbox failure (Liao et al., 2017). Continuous Wavelet Transform(CWT) obtained more accurate results than STFT in a study comparing the neutral current analysis performance of autotransformers (Aksenovich et al., 2020).

2.3. N-back Task

The N-back task is an artificial cognitive load task that remembers information before N steps from the last information when a series of information is presented, where N is the sequence of information that the subject should remember (Ranney et al., 2011). Cognitive load is a load in the cognitive process that occurs because the amount of information to be processed is greater than the amount of information that the brain can process (Paas et al., 2003). The N-back task blurs concentration and the value of N is proportional to the cognitive load (Jaeggi et al., 2010; Ranney, T. A. et al., 2011; Miller et al., 2009). Moreover, this approach, which can experimentally manipulate the level of working memory, was originally used to measure human short-term memory performance (Kirchner et al., 1958).

N-back task is often used for similar purposes in vehicle tests. Unni et al. (2017) performed multiple parallel tasks to measure the driver's working memory load level in a real scenario and measured the cognitive load with the N-back task as one of the tasks. Autonomous vehicle experiments were performed to confirm the safety of the vehicle or evaluate the driver's anxiety by controlling the driver's situational awareness (Harbluk et al., 2007; Ranney et al., 2011). He et al. (2019) changed the test according to the situation rather than the existing N-back task to apply cognitive load to the sense of sight, which is mainly used in driving situations. The number of two consecutive identical character pairs in one-back and the number of two duplicated identical character pairs in a string (may not be consecutive character pairs) in two-back were counted in this study. Various application methods for the N-back task have been developed and its use has been diversified (Jaeggi et al., 2010). Notably, our study pioneers the application of the N-back task to a personal mobility driving situation.

3. Research Framework

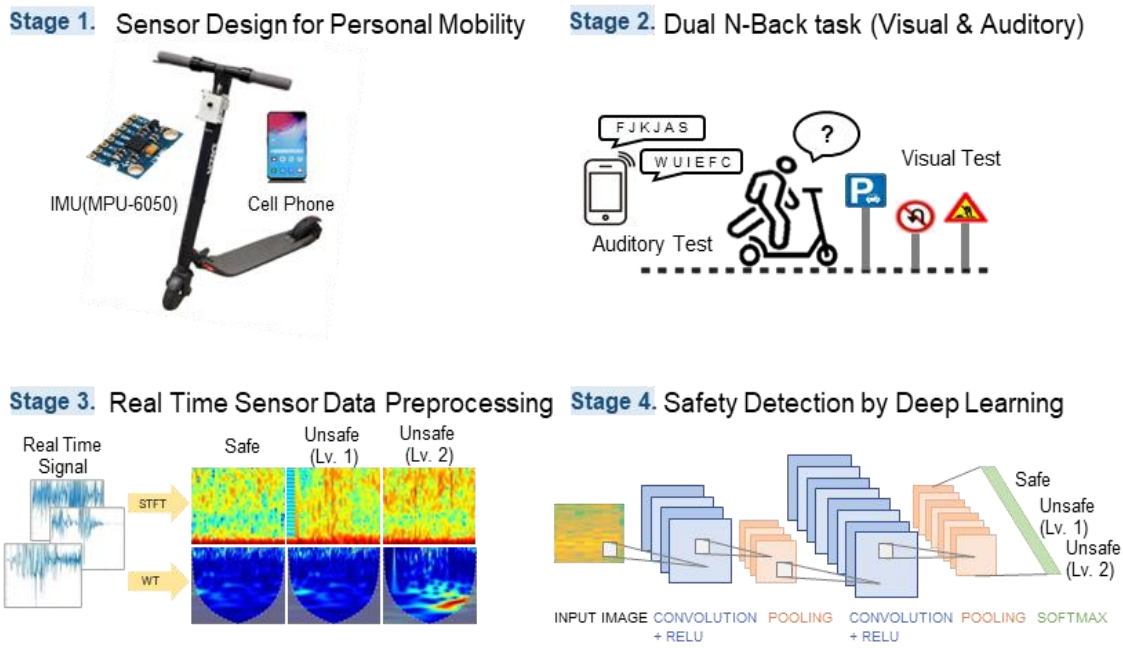

Figure 3: Research framework

The driver monitoring research framework presented in this study consists of four stages, as shown in Fig. 3. A Raspberry Pi device with a six-axis IMU sensor is attached near the handle of an e-scooter in Stage 1. The X-axis indicates up and down directions, the Y-axis refers to left and right directions, and the Z-axis demonstrates the direction of progress. The N-back Task experiment conducted in Stage 2 is divided into three sections of Safe, Unsafe (Lv.1), and Unsafe (Lv.2) within the experimental course. The driver goes through the course to produce cognitive loads on vision and hearing. STFT and WT techniques are used in Stage 3 for preprocessing and imaging vibration data which were stored in sensors when driving. Finally, a deep learning model that can classify three driving conditions using preprocessed image data is developed in Stage 4.

3.1 Stage 1: Sensor Design

The e-scooter used in the experiment is the Xiaomi Ninebot ES-2 model. Fig. 4 shows that the e-scooter is equipped with an IMU (MPU-6050) six-axis sensor connected to a Raspberry Pi and a Samsung Galaxy S7 device via a 3D-printed holder. The mobile phone and the IMU sensor are set up the same way. X-, Y-, and Z-axes of the accelerometer and gyroscope measure up and down, left and right, and front and back movements, respectively. The IMU sensors receive 100 Hz of data, but cellular phone sensors are not used in this study because they receive only 10 Hz of data. However, mobile phone holders were manufactured at the same time to ensure that sensor data, which will be used for future investigations, can be collected simultaneously.

A. 3D CAD design B. Actual holder C. Inside the holder

Figure 4: 3D-printed holder

3.2. Stage 2: N-back Test

3.2.1. Experimental Design

Seven subjects (women in their 20s who have ridden an e-scooter more than once) participated in this study. The experiment was conducted in downtown Seoul and lasted an average of 2 hours for two laps from Namyoung Station to Sinyeongsan Station. The course was divided into 15 sections, and data were obtained by presenting different situations for each section. Detailed intervals are shown in Table 1 and Fig. 5.

Safe-label sections 1, 2, 3, 9, and 10 consist of flat, uphill, and downhill terrains. These sections were selected to meet the actual situation and safety regulations. Unsafe (Lv.1)-label sections 4, 5, 6, and 7 present roadways (sections 4 and 6) and sidewalks (sections 5 and 7). Visual (sections 4 and 5) and auditory tests (sections 6 and 7) were conducted to create a situation where the driver fails to concentrate on driving due to cognitive load. Unsafe (Lv.2)-label sections 11, 12, 13, and 14 show equal roadways (sections 11 and 13) and sidewalks (sections 12 and 14). Although visual (sections 11 and 12) and auditory (sections 13 and 14) tests were also performed, experiments were conducted with a higher level of difficulty and stronger cognitive load than those of Lv.1. Sections 8 and 15 are construction sites; hence, no cognitive load test was carried out in these areas due to safety issues.

Fig. 5 shows the environment of the experimental driving section. Green lines on the map indicated the safe experimental sections without the N-back task (sections 1, 2, 3, 9, and 10). Blue lines denoted visual experiments during the N-back task (sections 4, 5, 11, and 12), and red lines refer to auditory experiments during the N-back task (sections 6, 7, 13, and 14). Two rounds of the corresponding course were completed, with easy Lv.1 N-back questions for the subjects in the first lap (sections 4, 5, 6, and 7) and difficult Lv.2 N-back questions for the subjects in the second lap (sections 11, 12, 13, and 14).

Section	Driving Road	Label	Sidewalks/Roadways
1	Namyeong Station–Samgakji Station	Safe (flat)	Bicycle path
2	Samgakji Station–Noksapyeong Station	Safe (uphill)	Bicycle path
3	Noksapyeong Station–Overpass	Safe (downhill)	Bicycle path
4	Overpass–Crosswalk in front of Han River Middle School	Unsafe Lv.1 (visual)	Roadways
5	Crosswalk in front of Han River Middle School–Seobinggo Station	Unsafe Lv.1 (visual)	Sidewalks
6	Seobinggo Station–Crosswalk	Unsafe Lv.1 (auditory)	Roadways
7	Crosswalk–National Museum of Korea	Unsafe Lv.1 (auditory)	Sidewalks
8	Construction site	-	Sidewalks
9	Samgakji Station–Noksapyeong Station	Safe (uphill)	Bicycle path
10	Noksapyeong Station–Overpass	Safe (downhill)	Bicycle path
11	Overpass–Crosswalk in front of Han River Middle School	Unsafe Lv.2 (visual)	Roadways
12	Crosswalk in front of Han River Middle School–Seobinggo Station	Unsafe Lv.2 (visual)	Sidewalks
13	Seobinggo Station–Crosswalk	Unsafe Lv.2 (auditory)	Roadways
14	Crosswalk–National Museum of Korea	Unsafe Lv.2 (auditory)	Sidewalks
15	Construction site	-	Sidewalks

Table 1: Label according to experimental interval

Figure 5: Experimental driving sections

3.2.2. N-back Task Design

We attempted to control the driver's situational awareness and intentionally lowered the concentration of driving by applying a cognitive load through the simultaneous progress of e-scooter driving and N-back task.

A specific number of listed alphabets were played via phone calls with wireless earphones while driving an e-scooter in the case of auditory tests. Participants then memorized the given set of letters one after another and matched all overlapping alphabets repeatedly during the driving period. Four letters were provided for the low-difficulty level and eight letters were given for the high-difficulty level for each problem. Thirty questions were asked per course and subjects were expected to answer in real time while driving.

Meanwhile, participants were asked to memorize road safety signs (Fig. 6) throughout the course in the correct order during the visual test. Twenty-four signs were placed along two driving courses, with twelve signs in each course. Only the color of the sign was memorized in the low-difficulty level, while the color, shape, and even some text information written on the sign were memorized in the high-difficulty level for each problem. The number of signs remained the same for each

difficulty level at 12 signs per driving course. If the subject successfully memorized the given tasks during the driving period, then three random questions were asked after completing each course.

A. Visual Test B. Road traffic signs used for visual testing

Figure 6: Photos of actual visual test driving and road signs used in the test

The degree of disturbance in driving concentration was asked for each course in a survey conducted on subjects after the experiment to ensure that the experiment was performed properly. The survey was scored using a five-point Likert scale, with 1 as the minimum disturbance and 5 as the maximum disturbance. As shown in Fig. 7, the average score was 1.86 (standard deviation of 0.86), 2.96 (standard deviation of 0.96), 3.68 points (standard deviation of 1.09) in the general safety, slightly anxious, and very anxious driving courses, respectively. It was demonstrated that the difficulty of the experiment increased with the increase of cognitive load. Furthermore, one-way ANOVA was performed on differences between groups, and the null hypothesis was rejected with a P value of 1.11E-06, thereby indicating the difference between groups.

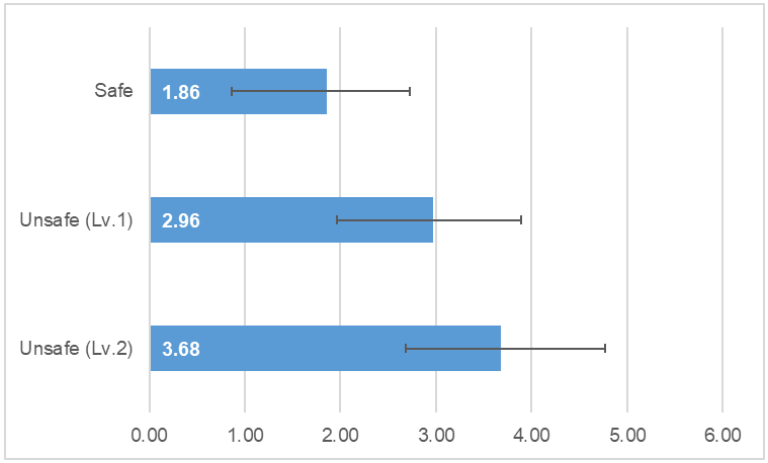

Figure 7: Degree of cognitive load for each experimental course

Analyzing the effects of this again according to hearing, vision, and driving environment (sidewalk/roadway), the difference in driving environment was insignificant. The side that chose the sidewalk cited reduced ride comfort primarily due to pavements and obstacles, while the side that selected the roadway cited the threat of surrounding vehicles as the main reason. However, Fig. 8 confirmed that more cognitive load is generated in the visual test with 3.50 (standard deviation of 0.84) and 3.14 (standard deviation of 1.24) points compared with that in the auditory test. T-test of

the two groups showed that the P value is 0.06, which is slightly higher than the significance level of 0.05.

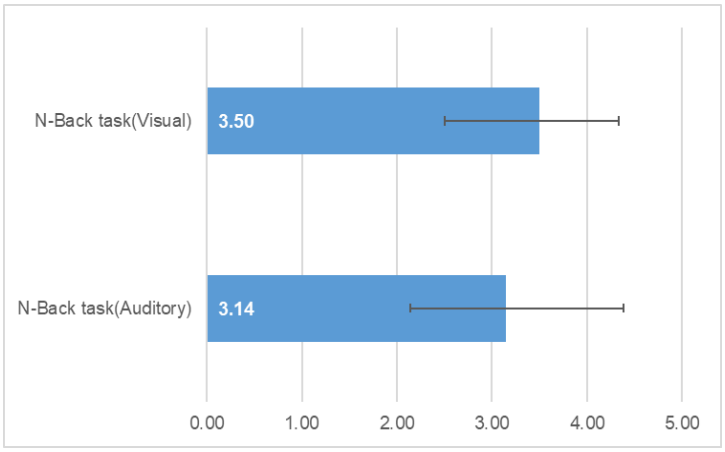

Figure 8: Degree of cognitive load for each visual/auditory course for the N-back task

3.3. Stage 3: Data Preprocessing

Vibration data (100 Hz) are received using the MPU-6050 IMU sensor attached to the Raspberry Pi device. Data are cut by 10 seconds for each course and then used. The time interval shifts every 1 second to ensure overlapping. A scalogram image (Fig. 9[a]) with a size of 224×224 was extracted with a window size of 40 in STFT using MATLAB (MathWorks, 2016).

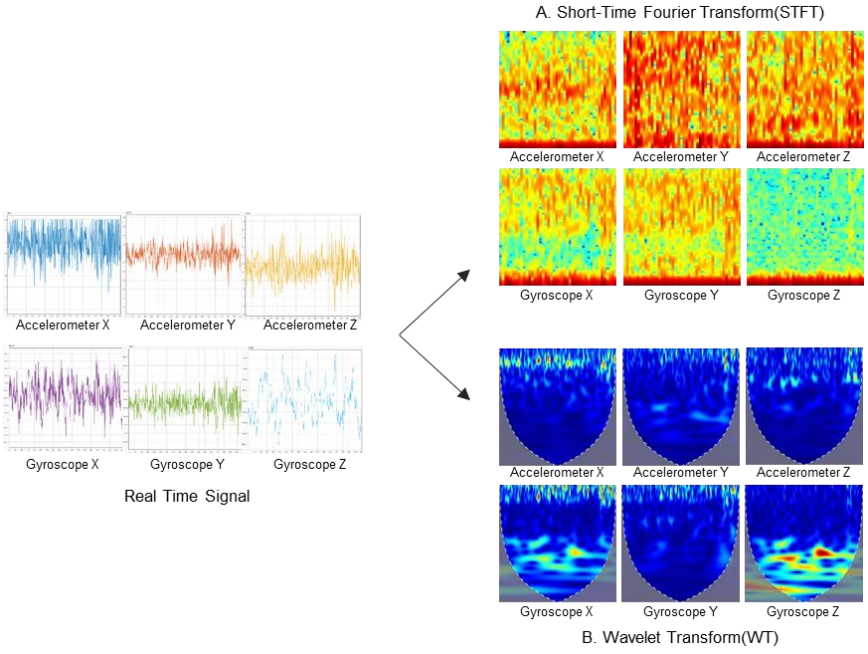

Figure 9: Image preprocessing of vibration data through signal conversion processing

A suitable window size is used for the frequency although the window size is not determined in the WT; hence, a scalogram with the same size as that in the STFT is extracted using MATLAB's CWT function. The scalogram obtained using the CWT function is shown in Fig. 9(b). The dashed white line in the figure contains the negative region from the edge of the line to the frequency or time axis and shows the areas of the scalogram that can potentially be affected by edge effects. Therefore, the information in the unshaded area within the dashed white line of the scalogram is an accurate time-frequency representation of data whereas that in the shaded area outside the dashed white line is less reliable due to the possibility of the edge effect.

3.4. Stage 4: Deep Learning

Input data are images with a size of 224×224 converted from six-sensor data into X-, Y-, and Z-axes of the accelerometer and gyroscope via STFT or WT. Learning data show an average of 1,700 images per driver, with 41% (approximately 690 images), 33% (approximately 560 images), and 26% (approximately 440 images) of labels for safety, unsafe Lv.1, and unsafe Lv.2 sections, respectively. The ratio of training and test sets for learning is 80:20.

The model follows the multimodal format, where individual deep learning models of six-sensor data images extract features in parallel and merge at the end to complement the information of different images on each axis and combine feature values to achieve increasingly accurate classification. These methods inform each other that a correlation exists when six different images are used as input data although unseen when individual models are trained with only one image (Ngiam, J. et al., 2011; Sohn, K. et al., 2014) and then integrate learned models in parallel to improve the robustness of predictions.

The CNN architecture used four models, namely, multimodal LeNet-5, multimodal VGG16, multimodal ResNet50 (Fig. 10), and multimodal DenseNet121 (Fig. 11). The three models, except for multimodal LeNet-5, used transfer learning. Transfer learning eases the problem of independent and identically distributed (i.i.d.) observations and solves the issue of insufficient data (Tan et al., 2018), thereby reducing the time and cost of collecting large amounts of data and rebuilding models (Pan, S. J., et al., 2009). We extracted the feature of vibration data using the weight of a pretrained model consisting of an existing ImageNet dataset as the initial value and subsequently classified it by adding two FC layers. Existing ImageNet models classify 1,000 different classes of data, but three outputs are designed in this study to pass the softmax layer to classify them into three. All models used Adam optimizer with a learning rate of 0.0001.

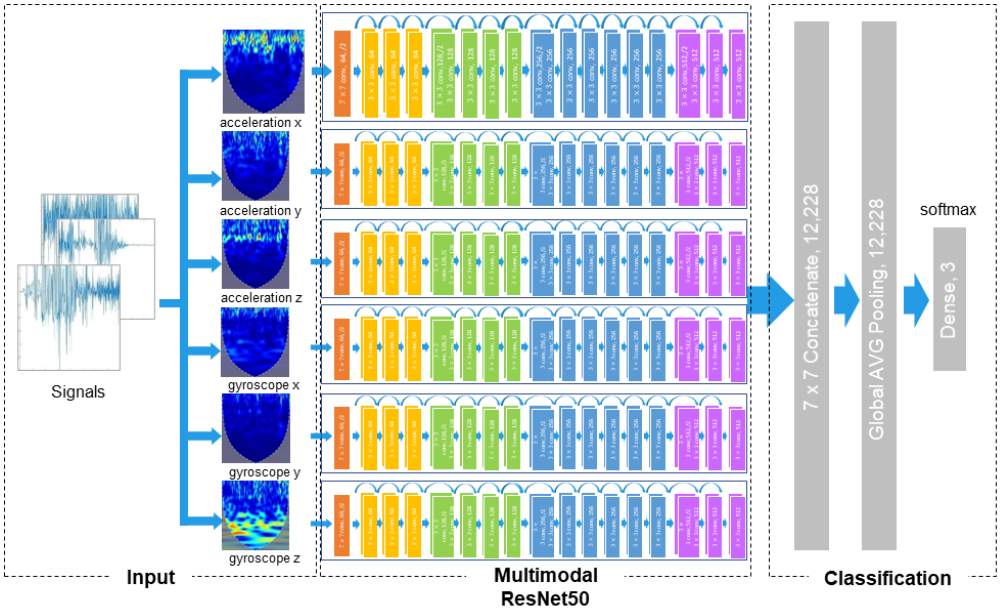

Figure 10: Multimodal ResNet50 network

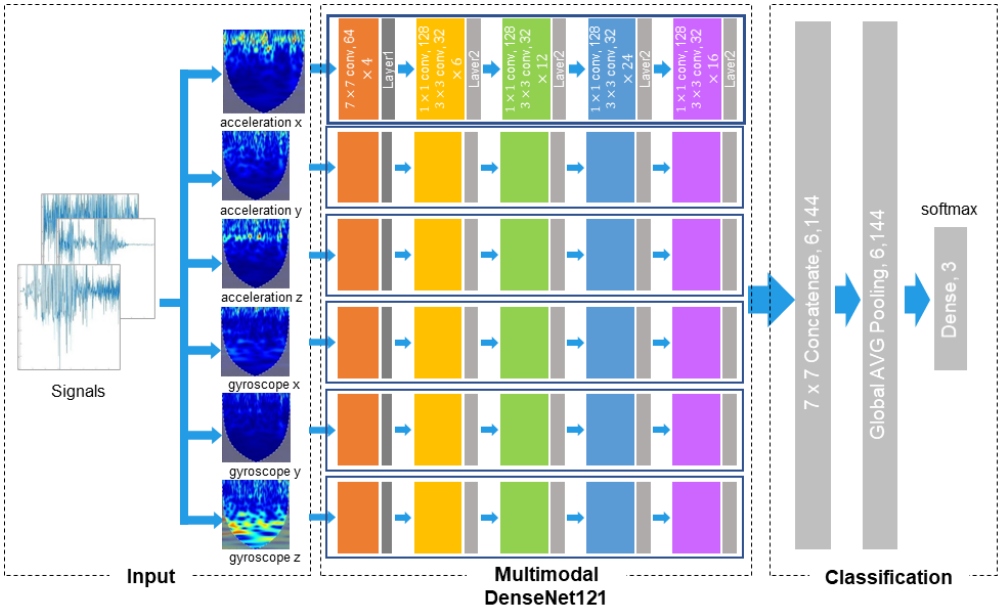

Figure 11: Multimodal DenseNet121 network

4. Experimental Results

4.1. Main Model Analysis

4.1.1. Model Comparison

The six images converted via STFT and WT (Section 3.3) are used as input data and trained using

the four models of multimodal LeNet-5, multimodal VGG16, multimodal ResNet50, and multimodal DenseNet121 (Section 3.4). The results of the individual model prediction of the seven drivers are listed in Table 2. Multimodal LeNet-5, a baseline shallow model, obtains an average accuracy of 43.36% (45.48%) and a standard deviation of 7.263% (4.669%) when images are converted using STFT (WT). Notably, Multimodal LeNet-5 is unsuitable for classifying the corresponding images because it obtained a large standard deviation and the lowest accuracy compared with the three other models.

The multimodal VGG16 model presents an average accuracy of 93.20% (97.06%) and a standard deviation of 6.510% (1.382%) when images are converted using STFT (WT). The accuracy of the VGG16 model, which shows a deeper architecture than the previous LeNet-5 model, significantly improves but continues to exhibit unstable results given its standard deviation.

The multimodal ResNet50 model demonstrates an average accuracy of 96.49% (96.58%) and a standard deviation of 0.017% (0.024%) when images are converted using STFT (WT). Compared with that of the VGG16 model, the accuracy of multimodal ResNet50 increases by approximately 3% when images are converted using STFT but slightly decreases by 0.5% when images are converted with WT. However, each model exhibits even accuracy when looking at the standard deviation.

Finally, multimodal DenseNet121, which presents the deepest layer among the four models, obtains an average accuracy of 99.10% (99.98%) and a standard deviation of 0.005% (0.008%) when images are converted using STFT (WT). This model obtained the best accuracy and standard deviation results compared with the three other models.

In conclusion, all four models showed better results with WT than STFT after data preprocessing and the CNN model showed that DenseNet121 achieves the highest accuracy. Moreover, sections are classified with high accuracy, as shown in the confusion matrix in Fig. 12, when the image preprocessed by WT is trained with the multimodal DenseNet121 model.

		Data preprocessing	
		STFT	WT
Deep learning architecture	Multimodal LeNet-5	43.36% (7.263%)*	45.48% (4.669%)
	Multimodal VGG16	93.20% (6.510%)	97.06% (1.382%)
	Multimodal ResNet50	96.49% (0.017%)	96.58% (0.024%)
	Multimodal DenseNet121	99.10% (0.005%)	99.98% (0.008%)

*The numbers in parentheses indicate the standard deviation.

Table 2: Accuracy and standard deviation of the deep learning model

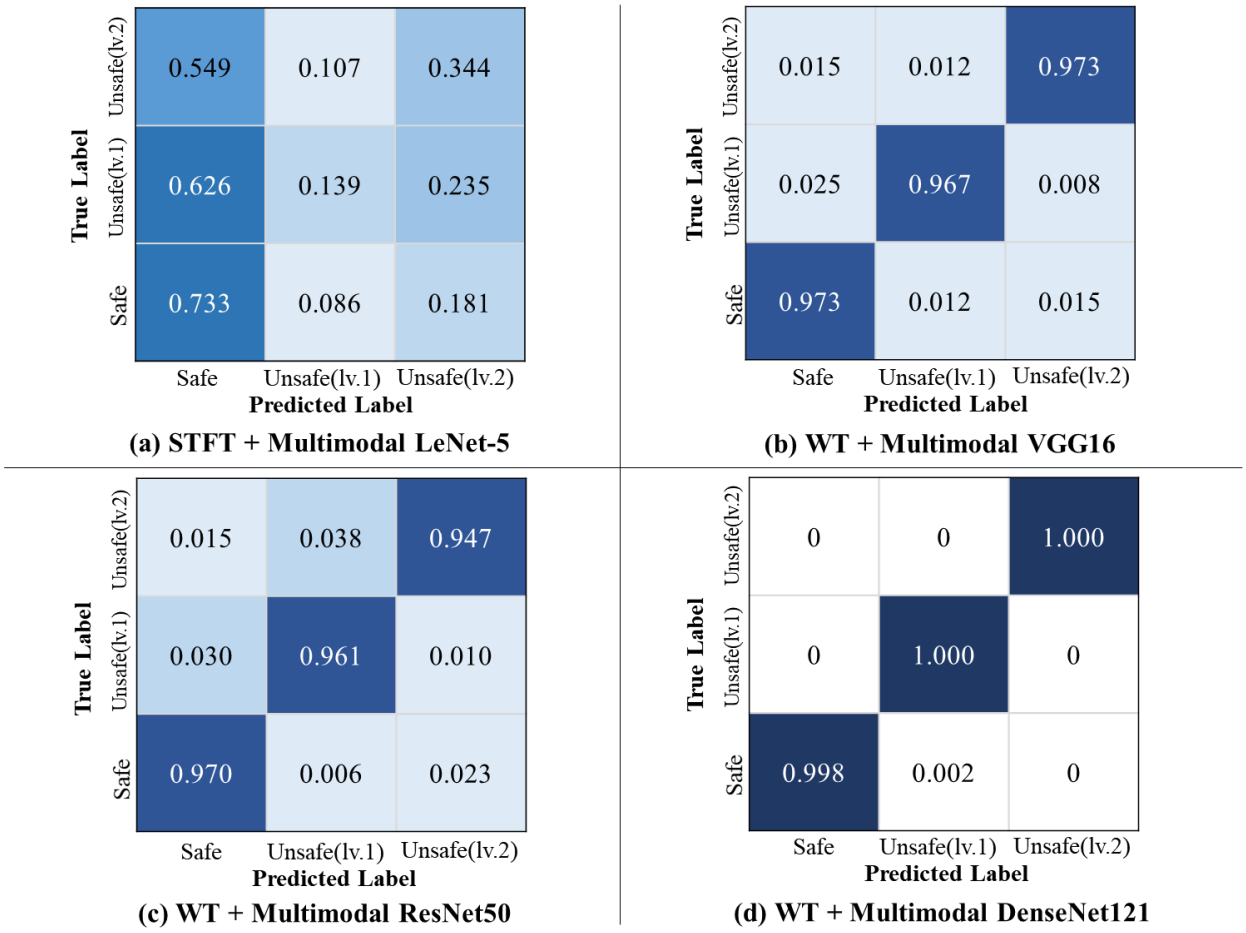

Figure 12: Confusion matrix of the deep learning model

4.1.2. Transfer Learning Effect

We analyzed transfer learning effect, and the results are listed in Table 3. The results of the three methods of multimodal VGG16, multimodal ResNet50, and multimodal DenseNet121, except for multimodal LeNet-5, were compared because only three architecture provide pre-trained models with ImageNet.

Three types of transfer learning were compared. First, training was performed using weights of the model trained with ImageNet data as initial values. Second, weights of the model trained with ImageNet data were fixed and only the FC layer was trained. Finally, training was conducted using only the architecture of the model and random initial values.

The multimodal VGG16 model demonstrates an accuracy of 97.06% and a standard deviation of 1.32% when weights trained with ImageNet data are used as initial values. The accuracy is 54.38%

and the standard deviation is 3.73% when the weight is fixed as the initial value. The accuracy is 55.48% and the standard deviation is 28.82% when a random initial value is used. It is shown that VGG16 is the architecture most affected by the pretrained weights.

The multimodal ResNet50 model presents an accuracy of 96.58% and a standard deviation of 0.02% when weights trained with ImageNet data are used as initial values. The accuracy is 44.06% and the standard deviation is 5.85% when the weight is fixed as the initial value. The accuracy is 98.82% and the standard deviation is 0.82% when an arbitrary initial value is used. Unlike the results of other models, the results of multimodal ResNet50 decreased by approximately 2% when the weight trained with ImageNet data was used as the initial value.

The Multimodal DenseNet121 model exhibits an accuracy of 99.98% and a standard deviation of 0.01% when weights trained with ImageNet data are used as initial values. The accuracy is 57.13% and the standard deviation is 3.47% when the weight is fixed as the initial value. The accuracy is 99.09% and the standard deviation is 0.76% when a random initial value is used. The Multimodal DenseNet121 model was not significantly affected by the pretrained weights compared to other models and had the highest accuracy for all types of transfer learning.

In conclusion, the multimodal VGG16 and multimodal DenseNet121 models showed the maximum accuracy when weights of pretrained model with ImageNet data were used as initial values. Meanwhile, all three models presented the minimum accuracy when pretrained weights were used without fine-tuning. Hence, it was found that when the pretrained weights from ImageNet are used as initial values, all weights have to be retrained using our data.

	Using pretrained weights as initial values	Freezing pretrained weights	Without using pretrained weights
Multimodal VGG16	97.06% (1.38%)	54.38% (3.73%)	55.48% (28.82%)
Multimodal ResNet50	96.58% (0.02%)	44.06% (5.85%)	98.82% (0.83%)
Multimodal DenseNet121	99.98% (0.01%)	57.13% (3.47%)	99.09% (0.76%)

Table 3: Comparison of transfer learning results

4.2. Additional Model Analysis

4.2.1. Single-modal Model

This study conducted multimodal deep learning with a six-axis sensor. In this section, we further

investigated which axis affected the accuracy given the difference in direction and information of each axis. The X-axis of the sensor attached to the e-scooter represents up and down directions, the Y-axis denotes left and right directions, and the Z-axis reflects front and back movements. We compared the performance of models trained on single-axis data as shown in Table 4. The single-modal model used multimodal DenseNet121 with the image converted via WT.

	X-axis	Y-axis	Z-axis
Accelerometer	99.73% (0.41%)	98.29% (1.52%)	98.31% (0.69%)
Gyroscope	97.83% (1.64%)	98.78% (1.03%)	96.83% (1.25%)

*The numbers in parentheses indicate the standard deviation.

Table 4: Accuracy and standard deviation for each axis of acceleration and gyroscope

The accelerometer X-axis showed the maximum accuracy with an average of 99.73% (standard deviation 0.41%). This finding is 0.25% less than the accuracy result of 99.98% when all six-axis sensors are used, thereby indicating that the cognitive load is highly related to the vibration of the e-scooter moving up and down. In addition, the overall accuracy of the accelerometer is higher than that of the gyroscope. The accelerometer accuracy was higher by 1.90% and 1.48% in the X- and Z-axes, respectively, but that in the Y-axis reduced to -0.49%.

4.2.2. Anomaly Detection

Autoencoder (AE) is a method widely used as a condition monitoring system when there is not enough abnormal data. AE restores the normal image without any difference when a normal image is used as the input but fails to restore properly when an abnormal image is used as the input. Therefore, anomaly detection can be performed by calculating the difference from the reconstructed image (Chalapathy et al., 2019).

In this section, further analysis was performed to determine if the AE could detect unsafe driving vibrations. Safe and unsafe data were denoted normal and abnormal images, respectively. We used convolutional autoencoder (CAE) with a loss function of mean squared error (MSE), and trained the model using only normal data. Abnormal data was used for testing purposes only.

In the test results, the reconstruction error of normal and abnormal images was 0.0222 (standard deviation of 0.0037) and 0.0220 (standard deviation 0.0040), respectively. There was no difference in the average reconstruction error between normal and abnormal data, the histogram in Fig. 14 illustrated that distribution of normal and abnormal are similar. Although the use of anomaly detection is beyond the scope of this study, we found that CAE is not suitable for the data we use.

However, as it is expensive to obtain abnormal data through experiments, further analysis on anomaly detection will be conducted in the future.

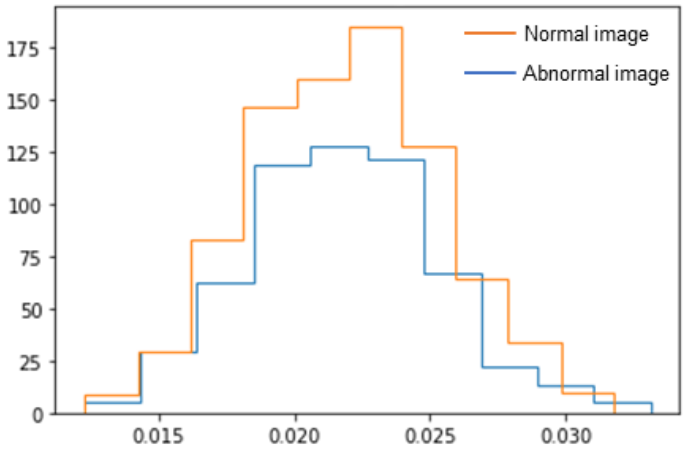

Figure 14: Distribution of reconstruction errors

5. Conclusion

This study proposes a personal mobility driver monitoring system to detect careless driving by using vibration data and deep learning. The N-back task is used to collect vibration data that occurs during unsafe driving. The vibration data are preprocessed into images using STFT and WT techniques and used to build multimodal deep learning models for classifying unsafe driving levels.

In the case of other companies, there were many studies and utilization of e-scooter using vision, but there were no studies to determine carelessness by converting the driver's driving type into vibration. This study is the first study to convert the anxiety of drivers on e-scooters into vibration data and classify them using deep learning. In addition, this is the first study to utilize N-Back Task, which was used only in simulations, by combining it with actual road driving to collect driver careless data. This study showed a high accuracy of more than 99% in prediction by exploring the preprocessing technique for converting vibration data into images and optimal deep learning model. The effectiveness of transfer learning was verified in the deep learning model, and the data importance of each axis was compared to show the possibility of a lightweight single model.

The contribution of this study is as follows. First, it is the first study to classify driver's careless driving using vibration data generated by e-scooters. Second, it is the first case of applying the N-back task to the e-scooter driving experiment. Third, we proposed a multimodal deep learning model with high prediction accuracy by exploring various signal preprocessing techniques and deep learning architectures. Lastly, the effect of transfer learning was verified, and a lightweight single-modal model was proposed as an alternative of multimodal model.

There are some limitations of this study. First, this study built only individual models due to lack of data. In future studies, we plan to increase the number of subjects. Then we will group the data by driving style and build a model for each group. Second, the vibration can be different depending on the location of the sensor attached to the e-scooter. In future research, we plan to find the optimal location by evaluating the sensitivity according to the location of the sensor attachment.

Acknowledgment

This study was supported by National Research Foundation of Korea (NRF) through grants funded by the Korean government (2017R1C1B2005266 and 2018R1A5A7025409).

References

Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M., & Inman, D. J. (2017). Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. Journal of Sound and Vibration, 388, 154-170.

Aksenovich, T. V. (2020, October). Comparison of the Use of Wavelet Transform and Short-Time Fourier Transform for the Study of Geomagnetically Induced Current in the Autotransformer Neutral. In 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon) (pp. 1-5). IEEE.

Berg Insight, (2020, April). The Bike and Scootersharing Telematics Market – 2nd Edition.

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407

Choi, How to safely drive an electric scooter? Technology development companies, The Hankyoreh, (Oct, 27, 2020), https://www.hani.co.kr/arti/economy/it/967429.html

Chong, Z. J., Qin, B., Bandyopadhyay, T., Wongpiromsarn, T., Rankin, E. S., Ang, M. H., ... & Low, K. H. (2011, September). Autonomous personal vehicle for the first-and last-mile transportation services. In 2011 IEEE 5th International Conference on Cybernetics and Intelligent Systems (CIS) (pp. 253-260). IEEE.

Devadasu, G., & Sushama, M. (2016, February). Identification of voltage quality problems under different types of Sag/Swell faults with Fast Fourier Transform analysis. In 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) (pp. 464-469). IEEE.

Facts & Factors, (2020, Fed). Personal Mobility Devices Market by Type (Wheelchair, Scooters, Handbikes, Walkers, Stair-lifts, Power Add on products and Others), and End User (Hospitals & Clinics, Ambulatory Surgical Centers, Urgent Care Center, Home Care Setting and Other End Users)– Global Industry Perspective Comprehensive Analysis and Forecast, 2019 – 2026

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., & Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. Accident Analysis & Prevention, 39(2), 372-379.

He, D., Donmez, B., Liu, C. C., & Plataniotis, K. N. (2019). High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified N-back task. IEEE Transactions on Human-Machine Systems, 49(4), 362-371.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. Memory, 18(4), 394-412.

Jurado, F., & Saenz, J. R. (2002). Comparison between discrete STFT and wavelets for the analysis of power quality events. Electric power systems research, 62(3), 183-190.

Khang, H. V., Karimi, H. R., & Robbersmyr, K. G. (2015, October). Bearing fault detection based on time-frequency representations of vibration signals. In 2015 18th International Conference on Electrical Machines and Systems (ICEMS) (pp. 1970-1975). IEEE.

Kendall Hutt, E-scooter injuries cost taxpayers nearly $15 million in two years, (Mar 07 2021). https://www.stuff.co.nz/national/health/124376214/escooter-injuries-cost-taxpayers-nearly-15-million-in-two-years

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. Journal of experimental psychology, 55(4), 352.

Korea Consumer Agency, Use the convenient electric scooter safely, (Mar, 21, 2019), https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=69445

Korea Transport Institute, Micro-mobility transportation policy support project, (June, 2017)https://www.koti.re.kr/component/file/ND_fileDownload.do?q_fileSn=106323&q_fileId=54c136ca-b641-4cd2-97fb-5d46181be27a

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Kwon, Electric scooter accidents have quadrupled in three years. Insurance in the future, The Chosunilbo, (Nov 07 2020). https://news.joins.com/article/23911294

LeCun, Y. (2015). LeNet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet, 20(5), 14.

Liao, Yixiao, Xueqiong Zeng, and Weihua Li. "Wavelet transform based convolutional neural network for gearbox fault classification." 2017 Prognostics and System Health Management Conference (PHM-Harbin). IEEE, 2017.

Mack, Electric Scooter Accidents Are Spiking Nationwide, Forbes (Jan 8 2020). https://www.forbes.com/sites/ericmack/2020/01/08/electric-scooter-injuries-are-spiking-nationwide/?sh=57d7d93f4698

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the n-back task a valid neuropsychological measure for assessing working memory. Archives of Clinical Neuropsychology, 24(7), 711-717.

Nell Lewis, E-scooters embrace AI to cut down on pedestrian collisions, CNN, (Feb 25 2021). https://edition.cnn.com/2021/02/24/business/e-scooter-safety-tech-ai-voi-spc-intl/index.html

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). Multimodal deep learning. In ICML.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. Educational psychologist, 38(1), 63-71.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

Ranney, T. A., Baldwin, G. H., Parmer, E., Domeyer, J., Martin, J., & Mazzae, E. N. (2011). Developing a test to measure distraction potential of in-vehicle information system tasks in production vehicles (No. HS-811 463).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sohn, K., Shang, W., & Lee, H. (2014). Improved multimodal deep learning with variation of information. Advances in neural information processing systems, 27, 2141-2149.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018, October). A survey on deep transfer learning. In International conference on artificial neural networks (pp. 270-279). Springer, Cham.

Thiruvaran, T., Phung, T., & Ambikairajah, E. (2013, April). Automatic identification of electric loads using switching transient current signals. In IEEE 2013 Tencon-Spring (pp. 252-256). IEEE.

Unni, A., Ihme, K., Jipp, M., & Rieger, J. W. (2017). Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: a realistic driving simulator study. Frontiers in human neuroscience, 11, 167.

Wang, L. H., Zhao, X. P., Wu, J. X., Xie, Y. Y., & Zhang, Y. H. (2017). Motor fault diagnosis based on short-time Fourier transform and convolutional neural network. Chinese Journal of Mechanical Engineering, 30(6), 1357-1368.

Wei, C., Chen, L. L., Song, Z. Z., Lou, X. G., & Li, D. D. (2020). EEG-based emotion recognition using simple recurrent units network and ensemble learning. Biomedical Signal Processing and Control, 58, 101756.