OXFORD CDE

Journal of Computational Design and Engineering, 2022, **9**, 1397–1409

DOI: 10.1093/jcde/qwac061
Advance access publication date: 26 July 2022
Research Article

Safety monitoring system of personal mobility driving using deep learning

Eunji Kim[1,3], Hanyoung Ryu[1], Hyunji Oh[1] and Namwoo Kang[2,3,*]

[1]Department of Mechanical Systems Engineering, Sookmyung Women's University, Seoul 04310, Republic of Korea
[2]Cho Chun Shik Graduate School of Mobility, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea
[3]Narnia Labs, Daejeon 34141, Republic of Korea
*Corresponding author. E-mail: nwkang@kaist.ac.kr
Note: A previous version of this manuscript was presented at the third Asia Pacific Conference of the Prognostics and Health Management Society (Jeju, Republic of Korea, Sep 8-11, 2021)

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 pre-processed using short-time Fourier transform and wavelet transform (WT) and then converted into an image (spectrogram). Third, four multimodal convolutional neural networks 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, wavelet transform, convolutional neural networks

1 Introduction

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 USA 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 USA, about 3300 patients were hospitalized for electric scooter-related injuries, a 365% increase. During the same period, all electric scooter-related injuries totalled 39 000, an increase of 222% (Mack, 2020). According to Stuff, New Zealand cost less than $15 million in taxes over 2 years due to electric scooter-related injuries. Between October 2018 and January 2021, a total of 6284 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 US 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 ultra-small 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 (Min-young, 2020).

Received: October 29, 2021. **Revised:** July 6, 2022. **Accepted:** July 7, 2022

This study aims to analyse 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. For example, if the driver does not stare at the front but sees an object or listens to music inducing a visual/auditory cognitive load, the electric scooter shakes unstably compared to when the driver concentrates on driving. 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 pre-processed image data.

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

2 Related Works

2.1 Spectrogram and CNN

In the era of Industry 4.0, the importance of human-machine interface (HMI) technology is ever increasing in various decision-making areas such as design and manufacturing (Gaudreault *et al.*, 2017; Zboinska, 2019; Scafà *et al.*, 2020). Our study, in particular, aims to design HMI for safety monitoring. There is previous research that studied safety monitoring with vibration data. There is a study that conducted safety monitoring for the surface using the pavement condition of the road (Alessandroni *et al.*, 2014; Zeng *et al.*, 2015). When collecting vibration data, low-cost sensors such as 3-axis accelerometers can be used to quickly identify road surface conditions (Cafiso *et al.*, 2020). A smartphone sensor can also be used to evaluate the roughness of the road surface using vertical acceleration (Yeganeh *et al.*, 2017). It can also collect vibration data from bicycles and electric scooters to detect road surface monitoring (Cafiso *et al.*, 2022).

These vibration and signal data are being actively used in deep learning (Matsushita *et al.*, 2021; Domala *et al.*, 2022). Our study focuses on converting such vibration data obtained from the sensor into a spectrogram image using STFT and WT methods. In a study that employed WT and CNN to diagnose gearbox failure, WT showed a more stable classification performance with less iterations than STFT (Liao *et al.*, 2017). Continuous wavelet transform(CWT) obtained more accurate results than STFT in a study comparing a neutral current analysis performance of autotransformers (Aksenovich, 2020).

These spectrogram images enable CNN-based deep learning. CNN 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 research. For example, two-dimensional (2D) CNN was used for gearbox failure signal detection by pre-processing 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 pre-processing (Abdeljaber *et al.*, 2017). In this study, we use and compare advanced CNN models such as LeNet-5 (LeCun, 2015), VGG16 (Simonyan & Zisserman, 2014), ResNet50 (He *et al.*, 2016), and DenseNet121 (Huang *et al.*, 2017) to classify normal and abnormal vibration data.

Recently, studies on STFT-CNN and WT-CNN have been used in various fields, showing high prediction accuracy. As a result of measuring the condition of transversely cracked road surfaces, manholes, and general road surfaces with an accelerometer mounted on a vehicle and a smartphone, the accuracy of WT-CNN and STFT-CNN were 97.2 and 91.4%, respectively (Chen *et al.*, 2021). STFT-CNN was used to detect the drone with the acoustic signal received by a microphone, showing an accuracy of 98.8% (Seo *et al.*, 2018). In a study on signal-to-noise related to communication, a detection probability of 90.2% was achieved as a result of using STFT-CNN (Chen *et al.*, 2020). A monitoring system of senescence effects on gait through step width was developed by analysing STFT-CNN and WT-CNN using Inertial Measurement Unit (IMU) sensors (Arshad *et al.*, 2021). For epileptic seizure detection using electroencephalogram signals, classification accuracy of 93.9 and 97.2% were achieved with STFT-CNN and STFT-LSTM (Beeraka *et al.*, 2022). STFT-CNN was also used to study information on pupil size and eye movement (Lee & Lee, 2018) and for the study of auscultation sound in the lungs due to Covid-19 which showed a high accuracy of 85.7% (Jung *et al.*, 2021).

2.2 *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 (Miller *et al.*, 2009; Jaeggi *et al.*, 2010; Ranney *et al.*, 2011). Moreover, this approach, which can experimentally manipulate the level of working memory, was originally used to measure human short-term memory performance (Kirchner,). Von Janczewski *et al.* (2021) investigates the effect of the *N*-back task on cognitive workload while driving, and shows the *N*-back task varies cognitive load substantially.

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

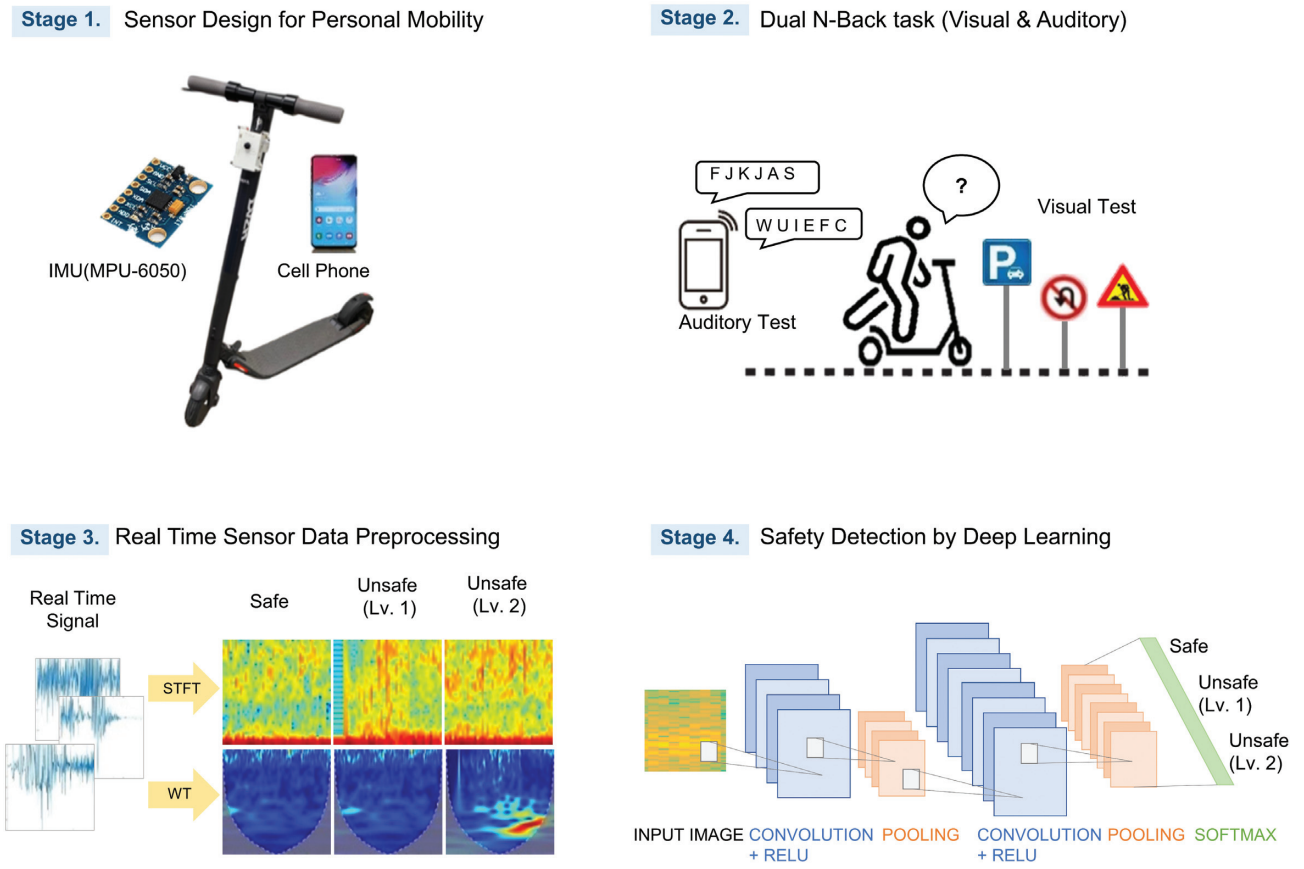

Figure 1: Experiment design.

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). Strayer *et al.* (2019) used the *N*-back task to interpret their results on cognitive workload when comparing smartphone-based systems (e.g. CarPlay and Android Auto) with native OEM (original equipment manufacturers) systems. Nilsson *et al.* (2020) studied car drivers' visual behaviour during execution of a cognitive task, and the cognitive load was generated by asking for an answer as quickly and accurately as possible. Notably, our study pioneers the application of the *N*-back task to a personal mobility driving situation.

3 Experiment Design

The driver monitoring experiment design presented in this study consists of four stages, as shown in Fig. 1. 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 pre-processing and imaging vibration data which were stored in sensors when driving. Finally, a deep learning model that can classify three driving conditions using pre-processed 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. Figure 2 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. The *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.

3.2 Stage 2: N-back test

i 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

Figure 2: 3D-printed holder.

Table 1: Label according to experimental interval.

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

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. 3.

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.

Figure 3 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).

ii 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. 4) throughout the course in the correct order during the visual test. 24 signs were placed along two driving courses, with 12 signs in each course. Only the colour of the sign was memorized in the low-difficulty level, while the colour, shape, and even some text information written on the sign were memorized in the high-difficulty level for each problem. The number of signs

Figure 3: Experimental driving sections.

Figure 4: Photos of actual visual test driving and road signs used in the test.

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.

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. 5, 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

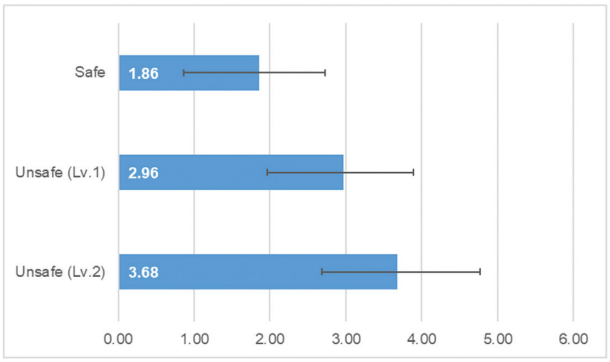

Figure 5: Degree of cognitive load for each experimental course.

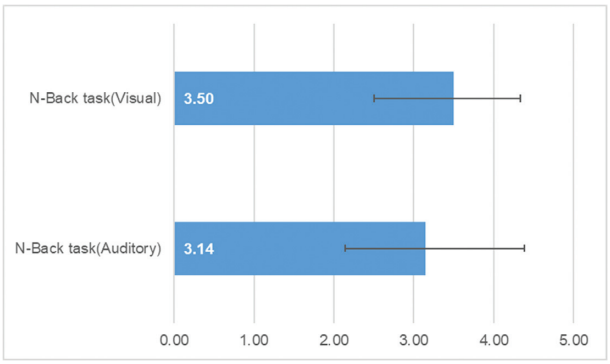

Figure 6: Degree of cognitive load for each visual/auditory course for the N-back task.

a *P*-value of 1.11E − 06, thereby indicating the difference between groups.

Analysing 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. 6 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.

3.3 Stage 3: data pre-processing

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. The overlapping rate of 90% was determined by parametric studies of deep learning results. A scalogram image (Fig. 7a) with a size of 224 × 224 was extracted with a window size of 40 in STFT using MATLAB (MathWorks, 2016).

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 through CWT function using predefined hyperparameter values in MATLAB. The scalogram obtained using the CWT function is shown in Fig. 7b. 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. Images for each level of WT and STFT are shown in Fig. 8. These images are for course 2 (Safe), course 5 (Unsafe (Lv.1)), and course 12 (Unsafe (Lv.2)) on the *x*-axis of acceleration. Using STFT and WT images alone, it was difficult to find a clear pattern that allows classification of the three classes without deep learning.

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 1700 images per driver, with 41% (approximately 690 images), 33% (approximately 560 images), and 26% (approximately 440 images) of labels for Safe, Unsafe Lv.1, and Unsafe Lv.2 sections, respectively. The training and test sets are sequentially split at a ratio of 80:20, avoiding overlap of data between training and test sets. The number of training and test set images corresponding to one axis for each individual driver is shown in Table 2

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 *et al.*, 2011; Sohn *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. 9), and multimodal DenseNet121 (Fig. 10). The three models, except for multimodal LeNet-5, used transfer learning. Transfer learning eases the problem of independent and identically distributed 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 & Yang, 2009). We extracted the feature of vibration data using the weight of a pre-trained model consisting of an existing ImageNet dataset as the initial value and subsequently classified it by adding two FC layers. Although existing ImageNet models classify 1000 different classes of data, our model classifies three classes as Safe, Unsafe (Lv.1), and Unsafe (Lv.2) through the softmax layer. All models used Adam optimizer with a learning rate of 0.0001. Four Titan XP GPUs with TensorFlow v.2.2.0 and Keras v.3.8 were used for training.

4 Experimental Results

4.1 Main model analysis

i 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 3. Multimodal LeNet-5, a baseline shallow model, obtains an average accuracy of 48.23% (55.64%) and a standard deviation of 7.41% (8.11%) when images are converted using STFT (WT). Notably, multimodal LeNet-5 is unsuitable for classifying the

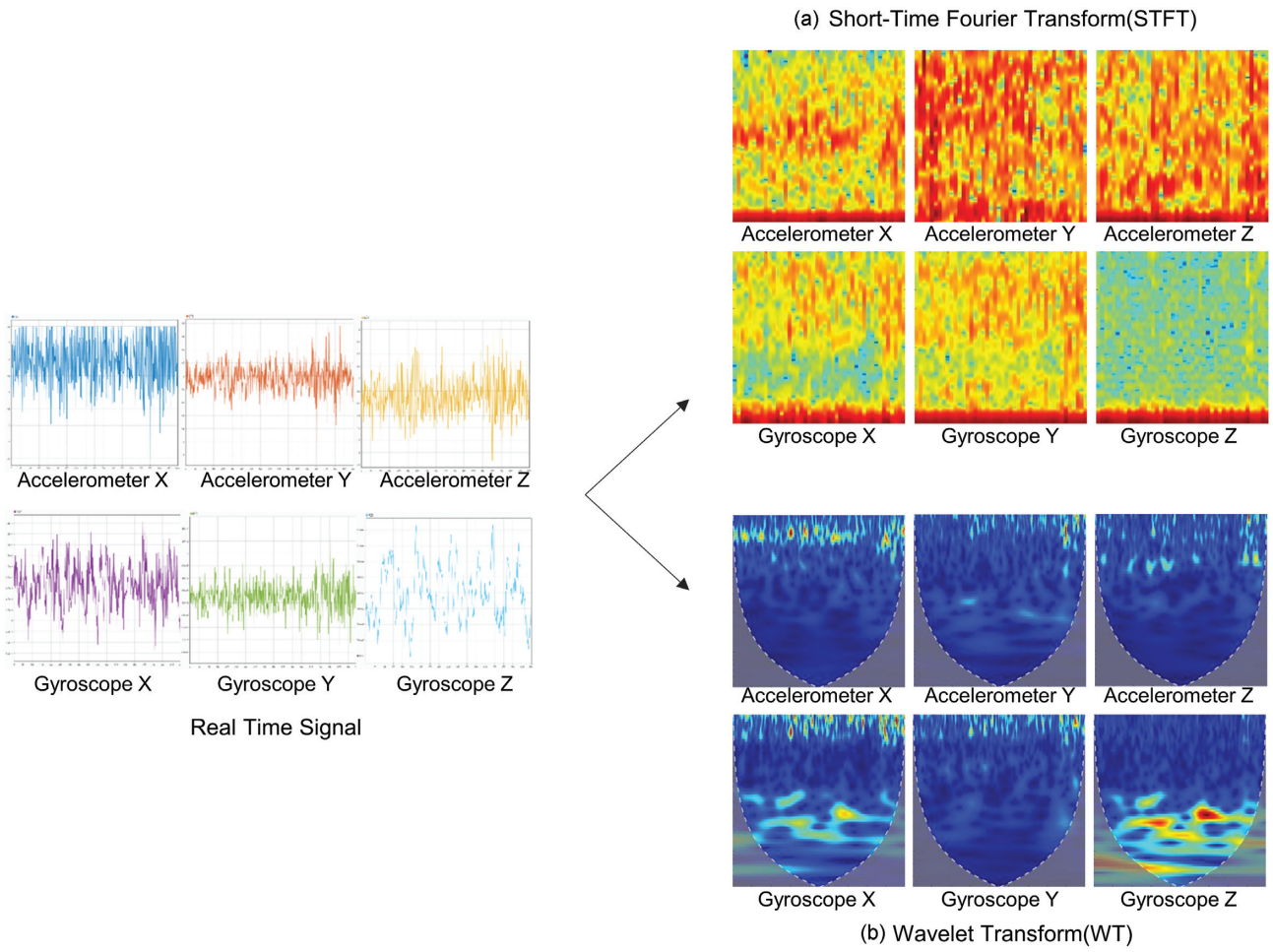

Figure 7: Image pre-processing of vibration data through signal conversion processing.

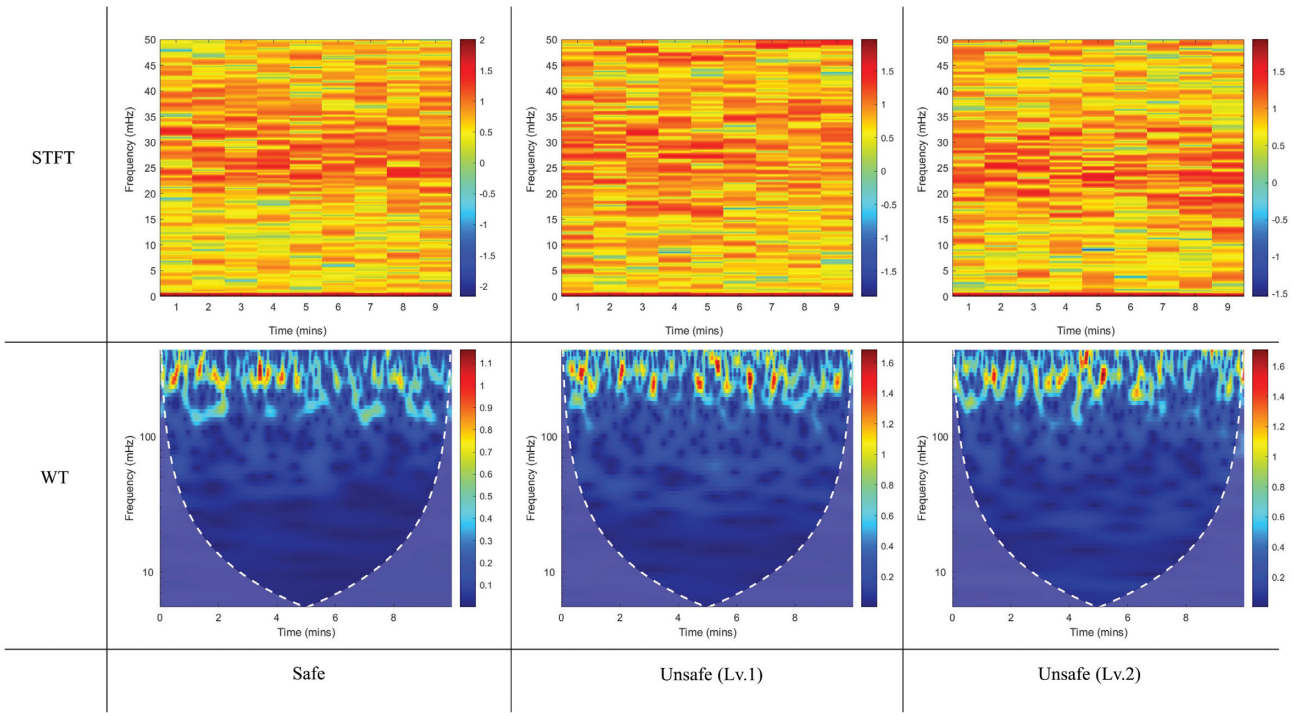

Figure 8: Pre-processed images for each class.

Table 2: Number of training and test set images for each driver.

Driver	Training set	Test set	Total
1	1066	268	1334
2	1559	392	1951
3	1758	440	2198
4	1754	440	2194
5	1134	285	1419
6	1472	359	1831
7	1296	326	1622

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 72.01% (86.02%) and a standard deviation of 4.25% (5.56%) 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 82.14% (66.10%) and a standard deviation of 5.21% (6.57%) when images are converted using STFT (WT). Compared with that of the VGG16 model, the accuracy of multimodal ResNet50 increases by approximately 10% when images are converted using STFT. However, it significantly decreases by 20% when images are converted with WT.

Finally, multimodal DenseNet121, which presents the deepest layer among the four models, obtains an average accuracy of 89.25% (91.82%) and a standard deviation of 2.04% (3.61%) 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, except for ResNet50 model, the other three models showed better results with WT than with STFT after data pre-processing 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. 11, when the image pre-processed by WT is trained with the multimodal DenseNet121 model.

ii Transfer learning effect

We analysed transfer learning effect, and the results are listed in Table 4. The results of the three methods of multimodal VGG16, multimodal ResNet50, and multimodal DenseNet121, except for multimodal LeNet-5, were compared because only three architectures provide pre-trained models with ImageNet. Two types of transfer learning were compared. First, training was performed using weights of the model trained with ImageNet data as initial values. Second, training was conducted using only the architecture of the model and random initial values.

The multimodal VGG16 model demonstrates an accuracy of 86.02% and a standard deviation of 5.56% when weights trained with ImageNet data are used as initial values. The accuracy is 42.62% and the standard deviation is 5.36% when a random initial value is used. It is shown that VGG16 is the architecture most affected by the pre-trained weights.

The multimodal ResNet50 model presents an accuracy of 66.10% and a standard deviation of 6.57% when weights trained with ImageNet data are used as initial values. The accuracy is 51.85% and the standard deviation is 8.02% when an arbitrary initial value is used. Like the results of other models, the accuracy of multimodal ResNet50 increased by approximately 15% when

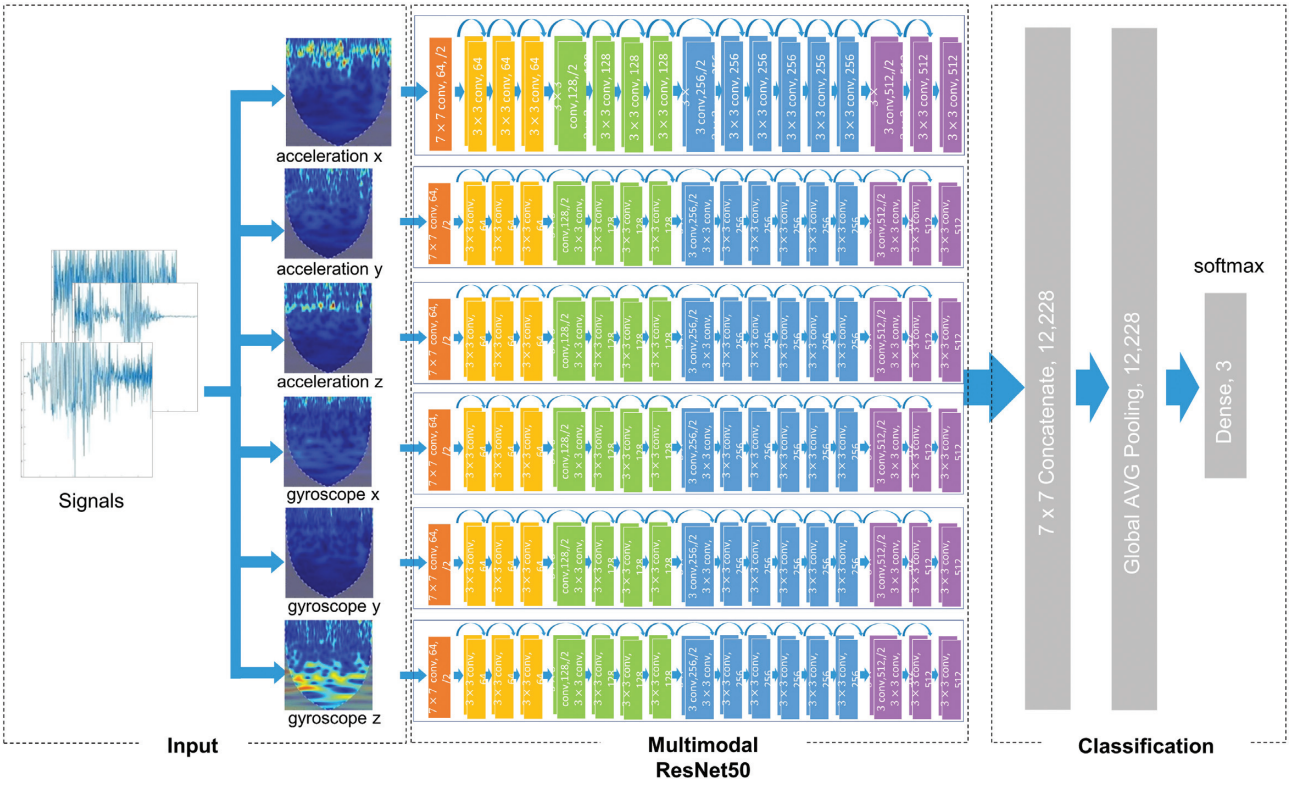

Figure 9: Multimodal ResNet50 network.

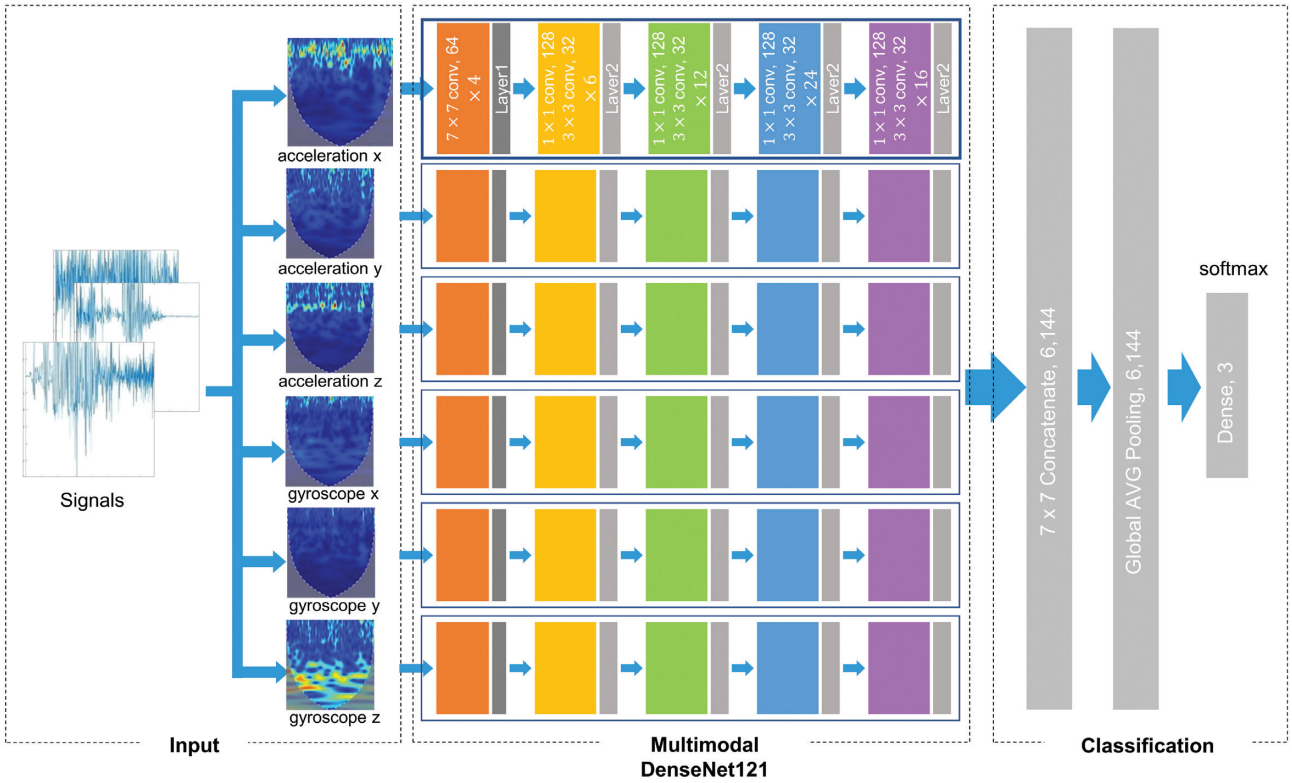

Figure 10: Multimodal DenseNet121 network.

Table 3: Accuracy and standard deviation of the deep learning model.

		Data pre-processing	
		STFT	WT
Deep learning architecture	Multimodal LeNet-5	48.23% (7.41%)*	55.64% (8.11%)
	Multimodal VGG16	72.01% (4.25%)	86.02% (5.56%)
	Multimodal ResNet50	82.14% (5.21%)	66.10% (6.57%)
	Multimodal DenseNet121	89.25% (2.04%)	91.82% (3.61%)

*The numbers in parentheses indicate the standard deviation. ($N = 7$).

the weights trained with ImageNet data was used as the initial value.

The multimodal DenseNet121 model exhibits an accuracy of 91.82% and a standard deviation of 3.61% when weights trained with ImageNet data are used as initial values. The accuracy is 53.26% and the standard deviation is 8.00% when a random initial value is used. The multimodal DenseNet121 model was significantly affected by the pre-trained weights compared to other model and had the highest accuracy for all types of transfer learning.

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

4.2 Additional model analysis

i 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 5. The single-modal model used multimodal DenseNet121 with the image converted via WT.

The accelerometer x-axis showed the maximum accuracy with an average of 88.58% (standard deviation of 1.58%). This finding is 3.24% less than the accuracy result of 91.82% 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

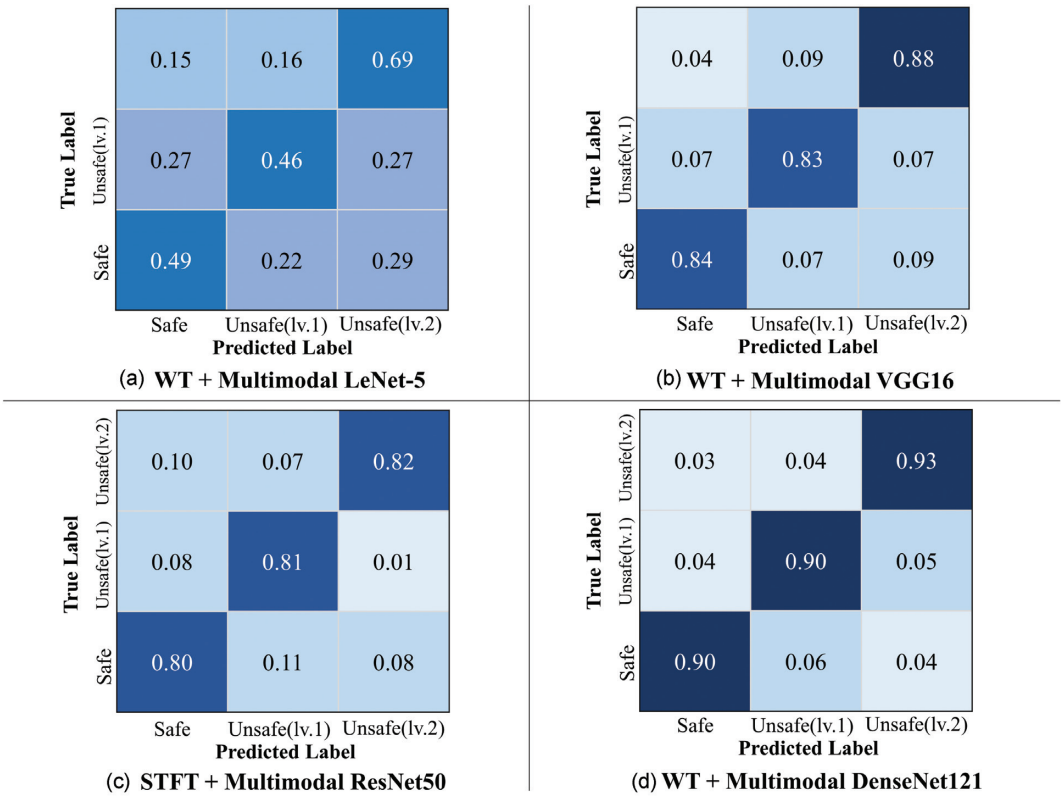

Figure 11: Confusion matrix of the deep learning model.

Table 4: Comparison of transfer learning results.

	Using pre-trained weights as initial values	Without using pre-trained weights
Multimodal	86.02%	42.62%
VGG16	(5.56%)	(5.36%)
Multimodal	66.10%	51.85%
ResNet50	(6.57%)	(8.02%)
Multimodal	91.82%	53.26%
DenseNet121	(3.61%)	(8.00%)

Table 5: Accuracy and standard deviation for each axis of acceleration and gyroscope.

	x-axis	y-axis	z-axis
Accelerometer	88.58%	87.19%	87.75%
	(1.58%)	(2.57%)	(3.40%)
Gyroscope	87.74%	87.38%	85.25%
	(3.28%)	(2.26%)	(4.21%)

*The numbers in parentheses indicate the standard deviation.

down. In addition, the overall accuracy of the accelerometer is higher than that of the gyroscope. The accelerometer accuracy was higher by 0.84 and 2.50% in the x- and z-axes, respectively, but that in the y-axis reduced to −0.19%.

ii *Fast Fourier transform analysis*

Frequency domain features were extracted from vibration data through fast Fourier transform (FFT). Table 6 shows the mean and standard deviation of the results obtained through FFT analysis of each individual driver. We used one-way ANOVA to check statistical significance of the difference between classes. The *P*-value of the MAX value was 0.32, the *P*-value of the MIN value was 0.62, the *P*-value of the root mean square (RMS) value was 0.18, and the *P*-value of the MEAN value was 0.57. As a result, there was no significant difference between classes. This shows that simple FFT analysis is not suitable for predicting inattentive driving.

iii *Heterogeneity*

In Section 4.1.1., the accuracy was evaluated with seven individual, independent models using training and test data of each driver. This was because the vibration patterns to cognitive loads are heterogeneous depending on each individual. Another approach is to test the driver's data not used for training with a model trained using other drivers' data. We trained a model with the data of six drivers and tested the model with the data of the remaining driver. The accuracy of the model was 43.07%, which is higher than the baseline value of 33.33% (three classes), but it still needs improvement. Currently, only seven drivers' data were used, but if grouping between drivers is possible through gathering more data, it is expected that the model from the same group could well predict a new driver's data.

Table 6: Results of FFT analysis.

Sensor	Level		MAX	MIN	RMS	MEAN
Accelerometer	Safe	AVG	9.33.E − 01	1.00.E − 04	1.64.E − 02	1.02.E − 02
		SD	1.32.E − 02	0.00.E + 00	1.90.E − 03	1.40.E − 03
	Unsafe (Lv.1)	AVG	8.96.E − 01	1.00.E − 04	1.61.E − 02	1.07.E − 02
		SD	5.70.E − 02	0.00.E + 00	1.60.E − 03	1.80.E − 03
	Unsafe (Lv.2)	AVG	9.18.E − 01	1.00.E − 04	1.82.E − 02	1.13.E − 02
		SD	2.03.E − 02	1.00.E − 04	1.20.E − 03	1.30.E − 03
Gyroscope	Safe	AVG	6.60.E + 03	3.70.E − 01	9.44.E + 01	4.08.E + 01
		SD	1.47.E + 01	1.33.E − 01	1.25.E + 01	9.44.E + 00
	Unsafe (Lv.1)	AVG	6.10.E + 03	3.51.E − 01	1.01.E + 02	5.01.E + 01
		SD	2.64.E + 01	1.10.E − 01	1.08.E + 01	1.01.E + 01
	Unsafe (Lv.2)	AVG	6.12.E + 03	4.59.E − 01	1.15.E + 02	5.44.E + 01
		SD	2.84.E + 01	1.32.E − 01	6.99.E + 00	5.91.E + 00

There may also be differences in the driving skills of the participants in the experiment, which may affect the model. When the result is divided into four high-skilled people and three low-skilled people, the models of the drivers with high driving skills does not well classify the levels of unsafe situation [Unsafe (Lv.1) vs. Unsafe (Lv.2)]. This is because the experienced driver is not significantly affected by the change in the magnitude of the cognitive load, although it does perceive the cognitive load itself. On the other hand, models of the drivers with low driving skills tended to accurately distinguish between the two unsafe classes. It can be seen that inexperienced drivers are sensitive to changes in the magnitude of cognitive load.

5 Conclusions

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 pre-processed 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 pre-processing 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 and novelty of this study are as follows. First, it is the first study to classify driver's careless driving using vibration data generated by e-scooters. This study shows that deep learning with spectrogram of vibration data can have high predictive performance in safety monitoring system of personal mobility. Second, it is the first case of applying the *N*-back task to the e-scooter driving experiment and deep learning. This study showed that inattentive driving of personal mobility can be simulated through the *N*-back task and that deep learning is a suitable method for training data from *N*-back task. Third, we proposed a multimodal deep learning model with high prediction accuracy by exploring various signal pre-processing 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, we plan to include individuals of diversity in age groups, sex, and ethnicity/race for the future research. Third, 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.

Acknowledgments

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

Conflict of interest statement

None declared.

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. https://doi.org/10.1016/j.jsv.2016.10.043.

Aksenovich, T. V. (2020). Comparison of the use of wavelet transform and short-time Fourier transform for the study of geomagnetically induced current in the autotransformer neutral. In *Proceedings of the 2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)* (pp. 1–5). IEEE. https://doi.org/10.1109/FarEastCon50210.2020.9271210.

Alessandroni, G., Klopfenstein, L. C., Delpriori, S., Dromedari, M., Luchetti, G., Paolini, B. D., Seraghiti, A., Lattanzi, E., Freschi, V., Carini, A., & Bogliolo, A. (2014). Smartroadsense: Collaborative road surface condition monitoring. In *Proceedings of the {UBICOMM} 2014: The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*(pp. 210–215). IARIA.

Arshad, M. Z., Jung, D., Park, M., Shin, H., Kim, J., & Mun, K. R. (2021). Gait-based frailty assessment using image representation of IMU

signals and deep CNN. In *Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1874–1879). IEEE. https://doi.org/10.1109/EMBC46164.2021.9630976.

Beeraka, S. M., Kumar, A., Sameer, M., Ghosh, S., & Gupta, B. (2022). Accuracy enhancement of epileptic seizure detection: A deep learning approach with hardware realization of STFT. *Circuits, Systems, and Signal Processing*, **41**, 461–484. https://doi.org/10.1007/s00034-021-01789-4.

Berg Insight. (2020). *The bike and scootersharing telematics market*. (2nd ed.). Berg Insight.

Cafiso, S., Di Graziano, A., Fedele, R., Marchetta, V., & Pratic`o, F. (2020). Sensor-based pavement diagnostic using acoustic signature for moduli estimation, *International Journal of Pavement Research and Technology*, **13**, 573–580. https://doi.org/10.1007/s42947-020-6007-4.

Cafiso, S., Di Graziano, A., Marchetta, V., & Pappalardo, G. (2022). Urban road pavements monitoring and assessment using bike and e-scooter as probe vehicles. *Case Studies in Construction Materials*, **16**, e00889. https://doi.org/10.1016/j.cscm.2022.e00889.

Chen, Z., Xu, Y. Q., Wang, H., & Guo, D. (2020). Deep STFT-CNN for spectrum sensing in cognitive radio. *IEEE Communications Letters*, **25**, 864–868. https://doi.org/10.1109/LCOMM.2020. 3037273.

Chen, C., Seo, H., & Zhao, Y. (2021). A novel pavement transverse cracks detection model using WT-CNN and STFT-CNN for smartphone data analysis. *International Journal of Pavement Engineering*, 1–13. https://doi.org/10.1080/10298436.2021.1945056.

Devadasu, G., & Sushama, M. (2016). Identification of voltage quality problems under different types of sag/swell faults with fast Fourier transform analysis. In *Proceedings of the 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)* (pp. 464–469). IEEE. https://doi.org/10.1109/AEEICB.2016.7538332.

Domala, V., Lee, W., & Kim, T. W. (2022). Wave data prediction with optimized machine learning and deep learning techniques. *Journal of Computational Design and Engineering*, **9**, 1107–1122. https://doi.org/10.1093/jcde/qwac048.

Facts & Factors. (2020). 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.

Gaudreault, J., Quimper, C. G., Marier, P., Bouchard, M., Chéné, F., & Bouchard, J. (2017). Designing a generic human-machine framework for real-time supply chain planning. *Journal of Computational Design and Engineering*, **4**, 69–85. https://doi.org/10.1016/j.jcde.2016.12.001.

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**, 372–379. https://doi.org/10.1016/j.aap.2006.08.013.

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). IEEE.

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**, 362–371.

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). IEEE. https://doi.org/10.1109/THMS.2019.2917194.

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

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**, 394–412. https://doi.org/10.1080/09658211003702171.

Jung, S. Y., Liao, C. H., Wu, Y. S., Yuan, S. M., & Sun, C. T. (2021). Efficiently classifying lung sounds through depthwise separable CNN models with fused STFT and MFCC features. *Diagnostics*, **11**, 732. https://doi.org/10.3390/diagnostics11040732.

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

Korea Transport Institute. (2017). Micro-mobility transportation policy support project. 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, S. Y. (2020). Electric scooter accidents have quadrupled in three years. Insurance in the future. The Chosunilbo. https://news.joins.com/article/23911294

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

Lee, H. J., & Lee, S. G. (2018). Arousal-valence recognition using CNN with STFT feature-combined image. *Electronics Letters*, **54**, 134–136. https://doi.org/10.1049/el.2017.3538.

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

Liao, Y., Zeng, X., & Li, W. (2017). Wavelet transform based convolutional neural network for gearbox fault classification. In *Proceedings of the 2017 Prognostics and System Health Management Conference (PHM-Harbin)* (pp. 1–6). IEEE. https://doi.org/10.1109/PHM.2017.8079274.

Mack, E. (2020). Electric scooter accidents are spiking nationwide. Forbes. https://www.forbes.com/sites/ericmack/2020/01/08/electric-scooter-injuries-are-spiking-nationwide/?sh=57d7d93f4698.

MathWorks(2016). CWT. https://kr.mathworks.com/help/wavelet/ref/cwt.html.

Matsushita, Y., Tran, D. T., Yamazoe, H., & Lee, J. H. (2021). Recent use of deep learning techniques in clinical applications based on gait: A survey. *Journal of Computational Design and Engineering*, **8**, 1499–1532. https://doi.org/10.1093/jcde/qwab054.

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**, 711–717. https://doi.org/10.1093/arclin/acp063.

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

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*. pp. 689–696.

Nilsson, E. J., Victor, T., Aust, M. L., Svanberg, B., Lindén, P., & Gustavsson, P. (2020). On-to-off-path gaze shift cancellations lead to gaze concentration in cognitively loaded car drivers: A simulator study exploring gaze patterns in relation to a cognitive task and the traffic environment. *Transportation Research Part F: Traffic Psychology and Behaviour*, **75**, 1–15. https://doi.org/10.1016/j.trf.2020.09.013.

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**, 63–71.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, **22**, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

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). National Highway Traffic Safety Administration.

Scafà, M., Marconi, M., & Germani, M. (2020). A critical review of symbiosis approaches in the context of Industry 4.0. *Journal of Computational Design and Engineering*, **7**, 269–278. https://doi.org/10.1093/jcde/qwaa022.

Seo, Y., Jang, B., & Im, S. (2018). Drone detection using convolutional neural networks with acoustic stft features. In *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*(pp. 1–6). IEEE. https://doi.org/10.1109/AVSS.2018.8639425.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/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.

Strayer, D. L., Cooper, J. M., McCarty, M. M., Getty, D. J., Wheatley, C. L., Motzkus, C. J., & Horrey, W. J. (2019). Visual and cognitive demands of carplay, android auto, and five native infotainment systems. *Human Factors*, **61**, 1371–1386. https://doi.org/10.1177/0018720819836575.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In *Proceedings of the International Conference on Artificial Neural Networks*(pp. 270–279). Springer, Cham. https://doi.org/10.1007/978-3-030-01424-7_27.

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. https://doi.org/10.3389/fnhum.2017.00167.

Von Janczewski, N., Wittmann, J., Engeln, A., Baumann, M., & Krauß, L. (2021). A meta-analysis of the N-back task while driving and its effects on cognitive workload. *Transportation Research Part F: Traffic Psychology and Behaviour*, **76**, 269–285. https://doi.org/10.1016/j.trf.2020.11.014.

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**, 1357–1368. https://doi.org/10.1007/s10033-017-0190-5.

Yeganeh, S. F., Mahmoudzadeh, A., Azizpour, M., & Golroo, A. (2017). Validation of smartphone-based pavement roughness measures. *AUT Journal of Civil Engineering*, **1**, 135–144. https://doi.org/10.22060/ajce.2017.13105.5328.

Zboinska, M. A. (2019). Influence of a hybrid digital toolset on the creative behaviors of designers in early-stage design. *Journal of Computational Design and Engineering*, **6**, 675–692. https://doi.org/10.1016/j.jcde.2018.12.002.

Zeng, H., Park, H., Fontaine, M. D., Smith, B. L., & McGhee, K. K. (2015) Identifying deficient pavement sections using an improved acceleration-based metric, *Transportation Research Record: Journal of the Transportation Research Board*, **2523**, 133–142. https://doi.org/10.3141/2523-15.