AUT Journal of Mathematics and Computing

AUT J. Math. Comput., 6(3) (2025) 257-268
https://doi.org/10.22060/AJMC.2023.22580.1177

Original Article

Driver cellphone usage detection using wavelet scattering and convolutional neural networks

Ali Besharati, Ali Nahvi*, Serajeddin Ebrahimian

Virtual Reality Laboratory, K.N. Toosi University of Technology, Tehran, Iran

ABSTRACT: This paper provides an automated system based on machine learning and computer vision to detect cellphone usage during driving. We used Wavelet Scattering Networks, which is a simple and efficient type of architecture. The presented model is straightforward and compact and requires little hyper-parameter tuning. The speed of this model is similar to the Convolutional Neural Networks. We monitored the driver from two viewpoints: a frontal view of the driver's face and a side view of the driver's whole body. We created a new dataset for the first viewpoint, and used a publicly available dataset for the second viewpoint. Our model achieved the test accuracy of 91% for our new dataset and 99% for the publicly available one.

Review History:
Received:25 July 2023
Revised:28 August 2023
Accepted:01 September 2023
Available Online:01 July 2025

Keywords:
Mobile use detection
Wavelet scattering network
CNN
Cascade object detector
Transfer learning

MSC (2020):
68T07; 68T45; 42C40

1. Introduction

Phone usage by drivers causes distraction and deviates attention from road hazards. About one-third of accidents caused by distraction involve a collision with a pedestrian, in which the driver had identified the pedestrian too late [46]. Accidents severity correlates with the type of inattention [45]– [56]. Inattention was the root of a third of severe accidents between 2011 and 2015 in Norway [7]. 80% of traffic accidents and 65% of close collisions are due to the distraction of the drivers for a period of 3 seconds before the accident [7]. On average, it takes 5 seconds to make a call, and if the car is moving at a speed of 100 km/h, it will travel 140 m during this time [5], so every pedestrian or vehicle in this distance is a potential cause of an accident. In Brazil, the fine for using mobile phones has increased by 150% in 5 years [40]. Using cell phones and texting while driving is responsible for 28% of accidents in the United States [16]. Every hour of the day, at least 5% of American drivers are using their cell phones while driving, which increases the risk of a crash 4-6 times [36], [31].

*Corresponding author.
E-mail addresses: ali_besharati@email.kntu.ac.ir, nahvi@kntu.ac.ir, sebrahimian@alumni.kntu.ac.ir

The technological capability for level-5 autonomous vehicles (AV) will be available by 2025, but the commercial availability will not happen until 2050 or even later [2]. This delay in availability has two reasons: first, the lack of substructions, like highway infrastructure [38] or 5G needed for AVs [19], and second, legislation challenges, like the legislative evolutions required in terms of administrative, civil, and criminal law [22]. Advanced Driver Assistance Systems (ADAS) are making progress in both of these spheres. Some examples of ADAS are compensating for the lack of attention for driving like cruise control [49], automatic car parking [12], and collision avoidance systems [18] or assisting the driver or correcting her/his behavior via warnings like drowsiness detection systems [11], [39] and drunkenness detection systems [41]. Such setups could be used for legislation or insurance related purposes [23]. The system proposed in this paper is an ADAS that could be used for warning, analyzing, and reporting drivers' behavior to authorities or insurance companies.

Distraction is defined as the deviation of the driver's focus to a marginal task instead of basic activities of driving [24]. Driver distraction can be divided into 4 categories: visual, auditory, manual, and cognitive [57]. Distraction caused by mobile phone includes all 4 types: looking at the phone screen (visual distraction), talking (auditory), texting (manual), and thinking about the discussed issues (cognitive) [57]. Sometimes drivers have compensatory beliefs [57] that a risky behavior is neutralized by performing a safe behavior. So the person thinks s/he can use the mobile phone because s/he drives at a low speed. Such a belief is dangerous or causes damage in many cases even at low speeds [57].

Several studies proposed methods for drivers' cellphone usage detection. Wagner et al [50] installed a head tracking camera (OptiTrack V120:Duo) in a test car on top of the dashboard's center console. They used 2 IR cameras (Flir FL3-U3- 13Y3M-C) with active IR illumination (two Osram SFH 4725S LEDs per side) to have a bright view of the driver's face. They claimed that their system remains fully functional during day and night light. They used a Raspberry Pi 3 Model B+ for image acquisition and brightness control. They presented 2 datasets for head position and driver posture. They made use of 4 different CNN architectures for classification using different cameras. He et al [20] established a dataset of 545 relevant images. For classification, they proposed the CornetNet-Lite network, which is based on the Hourglass module backbone to localize the cellphone in the image. They utilized 4 types of noise to measure the robustness of their model. Their training procedure included penalizing negative positions that are in a certain radius to positive positions. Rajput et al [34] extended the goal of "driver cellphone usage detection" to cellphone detection in all scenarios. For this purpose, they used the State Farm Distracted Driver Detection Kaggle dataset [27] as well as their own dataset named "IITH-dataset on mobile phone usage" (IITH-DMU), which has 618 images of 6000×4000 resolution. They annotated all images in both datasets and used Faster-RCNN and SSD with different backbones for transfer learning. The backbones they used were pre-trained on the COCO dataset. Likewise, Xiong et al [51] proposed a deep-learning-based approach. First, they used Progressive Calibration Networks [43] to detect and track the face and determine the candidate's mobile phone usage region. They used the YOLOV3 [35] network for the classification part. They also introduced their own experimental data, which was collected from 50 drivers under different lighting conditions through an infrared camera. It included 22216 calling behavior images and 25464 normal images. For the training of the neural network, they used the NVIDIA Jetson TX2 artificial intelligence supercomputer module. Behati et al [4] utilized the VGG16 architecture for this purpose. They used the Distracted Driver Dataset [1] to train their classifier. They also presented a shrunk version of VGG16, which is multiple times smaller than the original one but yields 1% less accuracy. Le et al [25] performed a robust deep-learning method to detect cellphone. They proposed a Multiple Scale Region-based CNN (RCNN) to detect the presence of a mobile phone near the face as well as the position of the hands and the steering wheel to ensure that both hands are placed on the steering wheel. As well as observing the steering wheel, they employed the same method to observe the face and localize the hand near the face. They used SHRP-2 Database and VIVA Hand Database for their experiment. Torres et al [48] also used a CNN trained on the first 4 classes of the State Farm dataset to perform the classification.

While recent studies in this field rely mostly on deep-learning-based approaches, most older studies employed classical methods for this classification. Xu et al [52] used a Near Infra-Red (NIR) imaging system described in [53] located on the roadside. Since the infrared light is absorbed by some vehicles windshield, they incorporated a powerful IR illuminator to compensate for the loss of light. They used Fisher vectors to categorize the driver's windshield as either cell phone usage violation or non-violation. Elqattan et al [13] used CNNs to diagnose this problem. They used UI-5240CP Rev.2 IDS camera with a zoom lens on the police cars and an Axis Q1645 camera on the roadside for this purpose. They made use of the YOLO algorithm trained on the COCO dataset to localize the car and the driver. To categorize the images, they cut and pre-processed each received image using contrast limited adaptive histogram equalization for enhancement and 3×3 Gaussian kernel for blurring. They utilized the transfer learning method to categorize the images. They used Xception [9] model pre-trained on ImageNet [10] dataset for transfer learning. Berri et al [6] proposed a hybrid system of pattern recognition and movement detection based on the fact that when drivers use their cell phones, they have a tendency to direct their gaze towards a point ahead, which limits the field of vision and affects the drivability. They implemented the detection using a camera attached to the dashboard of a car. First, their movement detection process classified the camera feed into 3 classes: "motionless", "cell phone to the ear", and "withdraw cell phone". Then, their pattern recognition process evaluated the image. The movement detection parametrized the pattern recognition by setting the classification threshold based on its output. They utilized a dataset of 100 positive images and 100 negative images. They used an MLP neural network with Gaussian and Sigmoid activation functions and applied a binary-coded genetic algorithm to find the best parameters for their network. Seshadri et al [42] used a supervised descent method (SDM) and Viola-Jones algorithm to track different points for investigating the desired area of the face, then the resulting images were given to the pre-trained classifier. For feature extraction, they utilized normalized raw pixels as well as features obtained from the histogram of oriented gradients. For the classification part, they used an ensemble of the Real Adaboost Classifier and SVM. Yaser et al [55] used a dataset of 49 positive and 30 negative images. Their images were taken from outside of a vehicle. They used cascade detectors for the classification part and reached the accuracy of 75%.

In this paper, first, we used a type of architecture known as Wavelet Scattering Network (WSN). Second, we showed that this network achieves higher accuracy like the widely used CNNs but using fewer parameters and relying only on strong feature extraction property of the wavelet transform. Moreover, it achieved high accuracy without applying image augmentation techniques to the images. Third, we produced a new dataset based on frontal view images of drivers using their cellphones.

The structure of this paper is as follows: In section 2, we provide a short description of the WSN and present the architecture used for this study. Also, we describe the datasets used for this study. In section 3, we explain the results of our method and compare the performance of our WSN with the widely used CNN backbones on the 2 datasets. In section 4, we compare our contribution with those of previous studies and we conclude the paper in section 5.

2. Materials and Methods

In this paper, we present a method based on machine learning to automatically detect driver cellphone usage. We created a dataset for this purpose and also used a publicly available dataset to improve the system's generalization. For classification, we used WSN and compared its results with the widely used CNN architectures. We show that our WSN architecture reaches the accuracy of the State of the Art (SOTA) models with its simple architecture and fewer number of parameters.

2.1. Ear Dataset

Like any other machine learning algorithm, our method requires a tremendous amount of data to learn the difference between "phone usage" and "safe behavior" [17]. First, we apply a HAAR cascade object detector [47] to the image to locate the subject's face and measure its size based on the distance from the camera. Then, using the face location and its size, we estimate the potential regions of the ears (Figure 1). There is a linear relationship between human face size and hand size [26]. So, after measuring the coefficients 0.4 and 1.2 for one subject, we can use it for other subjects. It enables us to robustly track the regions of interest (ROI)s, especially when the subject gets closer or farther to the camera.

2.2. Image acquisition protocol

We had the subjects sitting at a distance of 40-80cm from the camera. Hence, they held a cell phone in one hand beside the corresponding ear and gradually leaned to the left and right and also rotated their necks to the side while we took their images. We further asked each subject to repeat the same procedure with the same side of the face but using the opposite hand to have all possible cases in which a driver can use a phone while driving. The subjects were supposed to perform the moves gradually and we took 2 images from either side every second.———-Image acquisition protocol. We had the subjects sitting at a distance of 40-80cm from the camera. Hence, they held a cell phone in one hand beside the corresponding ear and gradually leaned to the left and right and also rotated their necks to the side while we took their images. We further asked each subject to repeat the same procedure with the same side of the face but using the opposite hand to have all possible cases in which a driver can use a phone while driving. The subjects were supposed to perform the moves gradually and we took 2 images from either side every second.

We only saved regions of ears to minimize memory usage and further processing. We took positive images from one hand and negative ones from the other. Due to the symmetry of the problem, there was no need to run the protocol on the other hand. Every image, whether positive or negative, could be flipped to represent that state for the opposite ear. This will be further discussed in the next subsection.

We gathered 18537 images from 48 subjects, 41 males and 7 females, using the protocol[1]. All images were taken with the subject's consent with a variety of indoor lighting setups.

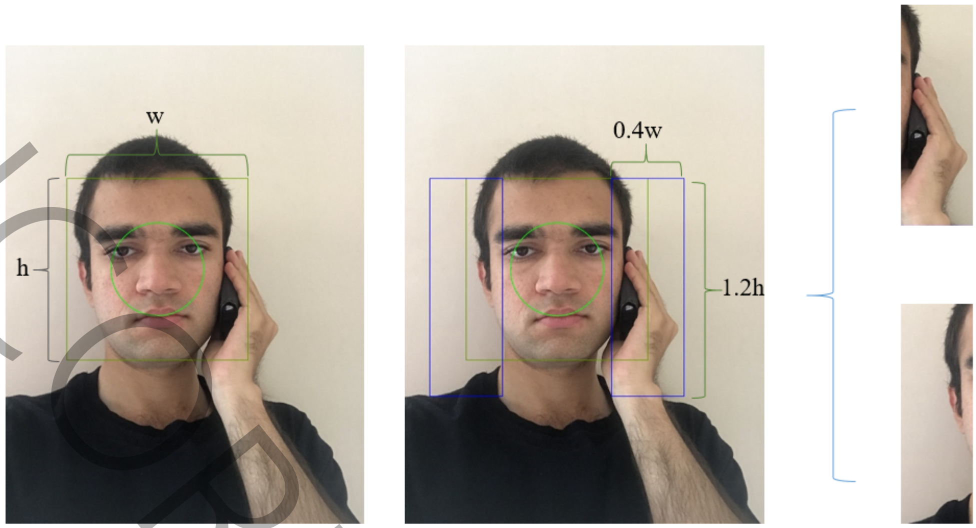

Figure 1: Left: Subject holding a phone in the left hand. The cascade detector locates the head (the circle) and measures the size (the square). Middle: We estimate the potential regions where the subject may have a phone based on the locations extracted by the cascade detector. Right: Extracted images for classifier decision

Image Augmentation. First, we resized all images to 150 × 50 (Figure 2). Then, we applied several image augmentation techniques to investigate the unsearched regions in the data space (Figure 3) [32]. Due to the horizontal symmetry of the human body, we applied the horizontal flip transform with a high probability of 0.9.

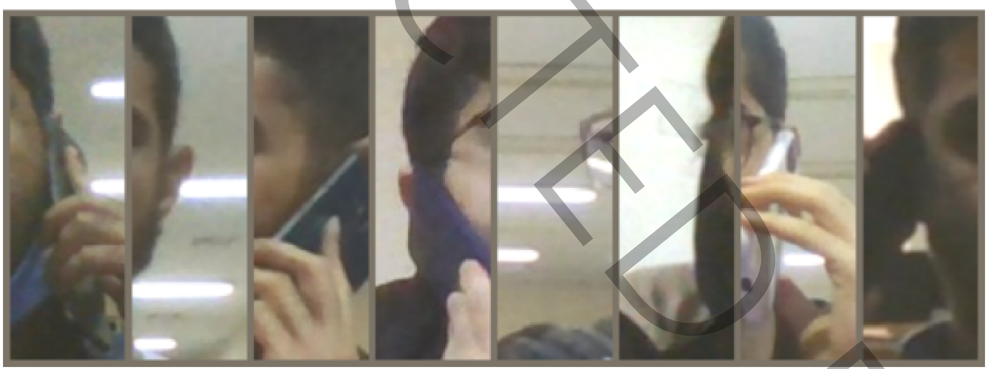

Figure 2: Raw images used for the validation set.

2.3. State Farm Datasets

We also used the first 5 classes of the publicly available Kaggle State Farm Dataset [27] (Figure 4). The total number of images used was 11745.

2.4. Classifier

We implemented a WSN [30] for each of the datasets. The WSN is a computationally efficient network based on wavelet filters that is invariant to translation and rotation [30]. The scattering transform requires 3 parameters for the transformation: number of samples or input size (T), the averaging scale (2^J), and the number of wavelets per octave, Q [3]. For our problem, T is the size of the images. We chose the numbers 3 and 8 for J and Q respectively. The scattering layer is a technique that involves several steps to encode important information

[1]The dataset is available at: https://www.kaggle.com/datasets/aliarch808/drivers-cellphone-usage-detection-via-ear-images

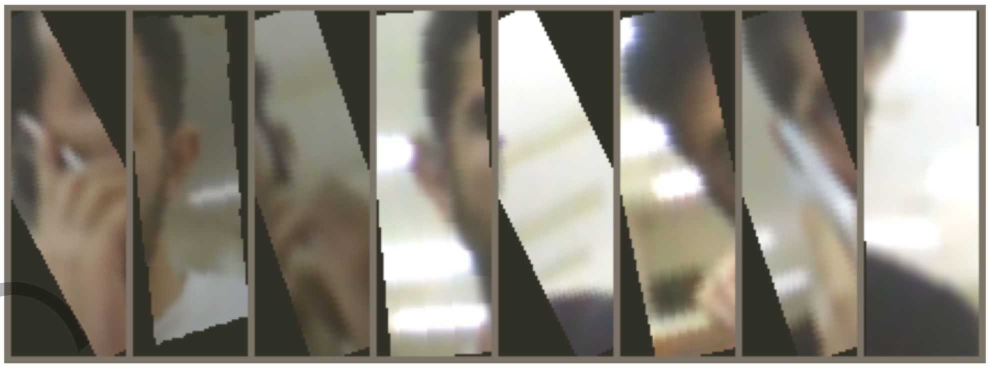

Figure 3: Transformed images used for the training set.

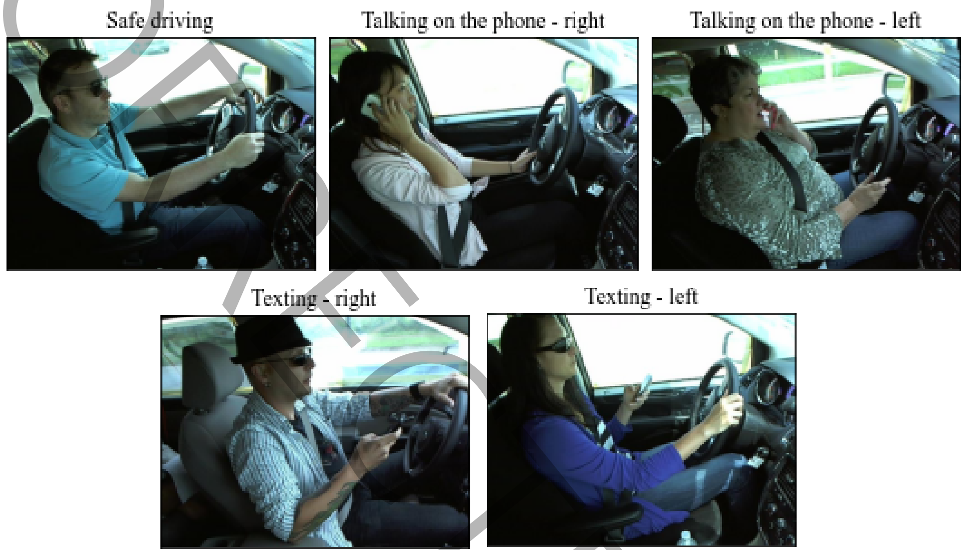

Figure 4: The first 5 classes of Kaggle State Farm Dataset [20]

about input images. The first step is the wavelet transform, which convolves the input image of the driver's face with a set of wavelet filters at different scales to capture facial features at different spatial frequencies and orientations. The wavelet coefficients are then processed to compute the scattering coefficients, which represent the result of cascading wavelet convolutions with non-linear modulus operations. After each wavelet convolution, the scattering layer performs a down-sampling operation to reduce the spatial dimensions of the coefficients and capture information at different scales. The layer then applies a non-linear modulus operation to the wavelet coefficients to capture amplitude information while discarding phase information, making the scattering transform translation-invariant. This process is repeated multiple times (3 times in this case, as specified by $J = 3$ in the layer definition), creating a cascade of transformations that capture increasingly higher-scale information while maintaining translation-invariance. Finally, the scattering layer outputs a set of scattering coefficients that encodes important information about the input images and pass them to the next layer for further processing [30], [3]. Figure 5 depicts the architectures we used for either of the datasets.

We also used three CNNs to have a comparison amongst different architectures. We claim that our WSN achieves the same accuracy as SOTA architectures but with far fewer parameters.

We resized all images in Ear Dataset to 150x50 and all images in State Farm dataset to 64x64. We used grayscale images for the WSN and RGB augmented images for the CNNs. Both datasets were split in 2 (90% of the data was used for training and 10% for validation) and we shuffled the data during training to avoid overfitting. The first model we used for this purpose was a baseline CNN with 7 layers and 620K and 1.28M parameters for the State Farm and the Ear dataset, respectively. The other two were VGG16 [44] and ResNet50 [21], which are two well-known architectures widely used in machine vision tasks. We trained all networks for 50 epochs with Adam optimizer.

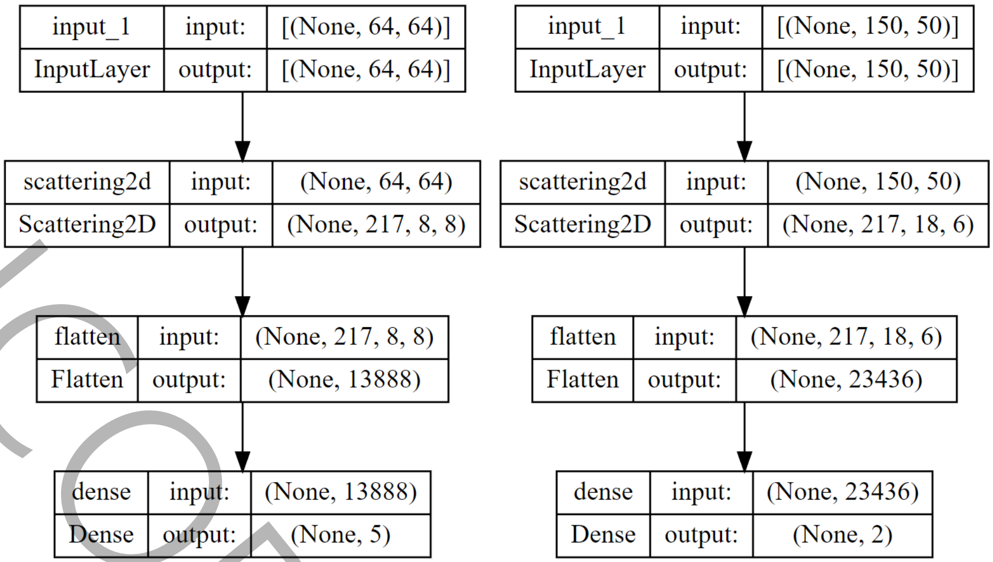

Figure 5: WSN Architectures used for the classification of: Left: State Farm data classifier and Right: Ear data Classifier

2.5. Tools

All the codes were written in Python. The fundamental libraries were OpenCV for image processing tasks and PyTorch, Tensorflow, and Keras for machine learning tasks [15].

All of the trainings of the Neural Networks were done using Google CoLaboratory™ data analysis tool and its Tesla T4 GPU. For inferencing, we used a laptop computer with an Intel Core i7 CPU, 8 GB RAM, and a webcam.

3. Results

In this section, we provide the results of our models both in the training and the inference phase:

3.1. Training

All the CNNs we used reached 99% accuracy. Figures 6 and 7 compare the accuracy and loss of the CNNs with the WSN for the Ear dataset and State Farm dataset respectively for training for 50 epochs. The high validation accuracy reached by the WSN shows that it surmounted overfit to a high degree. For the Ear dataset, due to high similarity between classes, it falls behind the CNNs in terms of validation accuracy but for the State Farm dataset, it accomplishes the same validation accuracy. The fluctuations of models in loss and accuracy plots could be attributed to two factors. First, the high similarity between the images of the two classes of the ear dataset makes the learning challenging and creates temporary wrong directions for the network. Secondly, the more prominent reason for the fluctuations could be attributed to the small number of parameters in the SWN models. Due to this sparsity of the parameters, the networks are forced to make their weights more fluid, and result in high fluctuations in the loss and accuracy metrics.

3.2. Inferencing

Table 1 shows the correctness metrics of SWN model on each of the Ear and State Farm datasets.

Table 1: SWN Correction Metrics for each of the datasets

Dataset	% Accuracy	% Precision	% Sensitivity	% Specificity
Ear	91	87	94	89
State Farm	99	99	99	99

Table 2 provides the utility metrics for the models trained on Ear Dataset and State Farm Dataset. It provides a comparison between the WSN and the CNNs in terms of volume and speed and shows that WSN is many times smaller than the other networks.

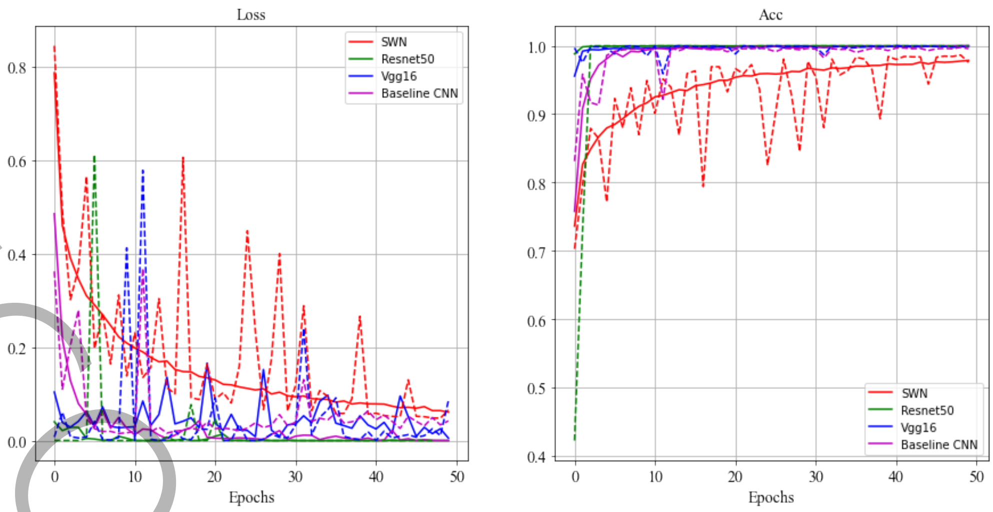

Figure 6: Loss and Accuracy of WSN, Resnet50, VGG16 and the baseline CNN for the Ear dataset (Dashed lines show the validation loss and accuracy)

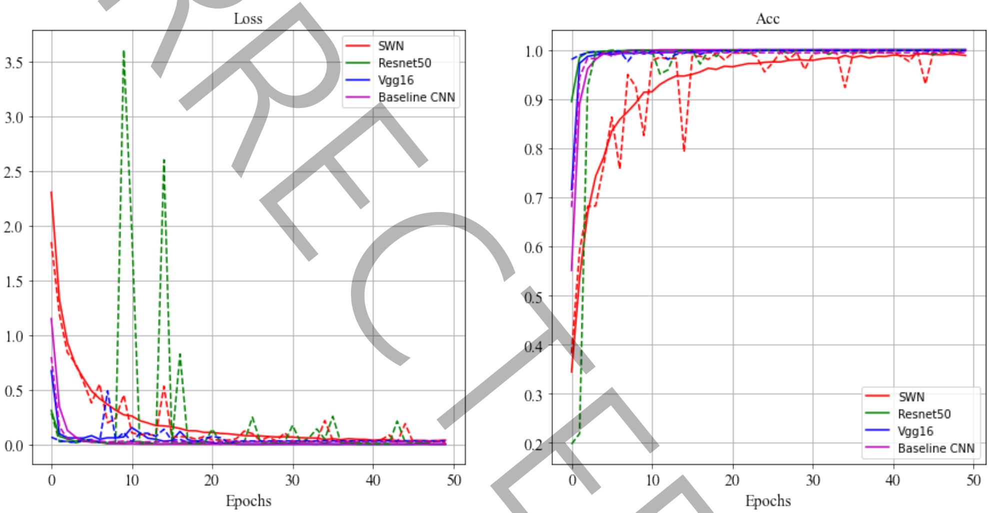

Figure 7: Loss and Accuracy of WSN, Resnet50, VGG16 and the baseline CNN for the State Farm dataset (Dashed lines show the validation loss and accuracy)

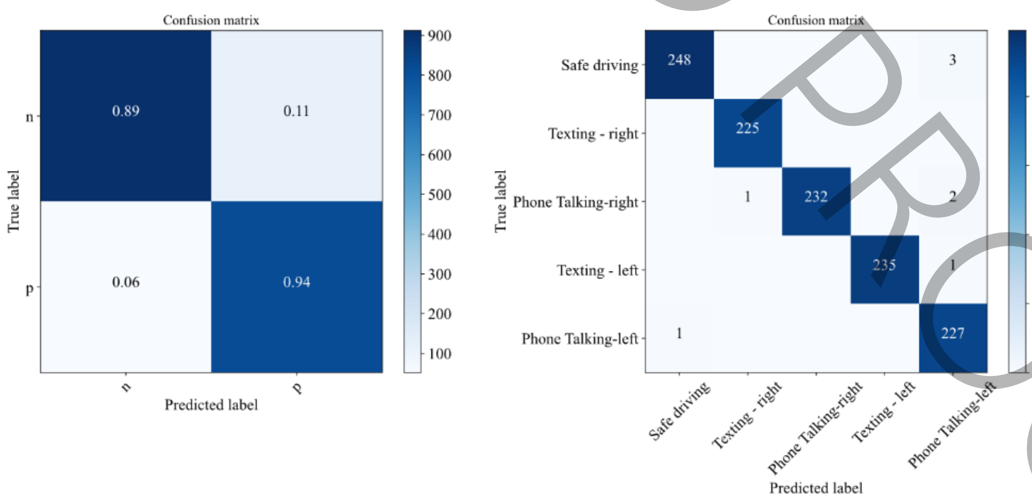

Figure 8: Confusion matrices for the classification of Ear dataset (Left) and the State Farm dataset (Right)

Table 2: Utility metrics

	Ear Dataset		State Farm Dataset	
Model	fps	Volume (MB)	fps	Volume (MB)
WSN	5	**0.57**	5	**0.84**
VGG16	3	174.9	7	179.1
Resnet50	8	330.7	8	301.8

4. Discussions

In this section, we elaborate on our results and compare our method with previous studies. Table 3 compares the performance of this paper with other research works. It indicates that this study achieved the correctness metrics of the previous studies.

Table 4 compares the backbones used in other studies. It shows that the number of parameters in WSN is 300 times less than the smallest model used by previous authors (ResNeXt-50 in [50]) for this classification. In addition, it is about 12 times smaller in volume. Most recent studies tried to tackle the problem using CNNs and achieved SOTA results. Our method gained the same accuracy, but with far fewer parameters. Additionally, as illustrated in figure 5, its 2-layer architecture is a lot simpler than the CNNs and does not require pooling or drop out layers. As indicated in figures 7 and 8, small differences between validation and training graphs show that the proposed architecture and hyperparameters were suitable for the datasets without overfitting.

Table 3: Correctness metrics and fps for other studies and this study

Study	% Accuracy	% Precision	% Sensitivity	fps
Wagner et al. 2022	92.8 & 90	–	–	44 & 28
He et al. 2021	86.2	81.4 & 86.2	75.3 & 83.8	33
Rajput et el. 2020	98 & 95	–	57.5 & 82.4	–
Xiong et al. 2019	96	–	–	25
Elqattan et al. 2019	89 & 95	46	100	–
Baheti et al. 2018	95.5	–	–	–
This study	91 & 99	87 & 99	94 & 99	5 & 5

Table 4: The volume and the number of parameters of the backbone architectures for previous studies and this study

Study	Model Name	Model Volume	Number of Parameters
Wagner et al. 2022	ResNeXt-50	10.84 MB	23 M
He et al. 2021	CornerNet-Lite	$\sim$ 600 MB	–
Rajput et el. 2020	Inception-v2	92 MB	23.9 M
Xiong et al. 2019 & Elqattan et al. 2019	YOLOV3	235 MB	62.25 M
Baheti et al. 2018	VGG16	528 MB	138.4 M
This study	WSN	**0.84MB**	**69.4 k**

As explained in section 2.3, we used RGB augmented images for CNNs training, but we trained the WSNs with grayscale images without any augmentation. This highlights the high potential of the WSN in feature extraction. From this point of view and with a consideration of the results achieved for the State Farm dataset, we observe that the WSN accomplishes SOTA accuracy by training on one-third of the information used for the CNNs (grayscale images versus RGB images).

The main challenge we faced in this study was the weaknesses of the HAAR cascade detector that affected the performance of the classifier for the frontal view. Face occlusions or dark skin harms the functionality of the cascade detector, hence we would lose the track of the face and the ears' ROIs. In this regard, we refer to the studies that worked on occluded face detection [8], [33]. The error rate of our system increases if the environment is too dark. Also, we cannot detect cellphone usage if the subject uses a hands-free gadget for the phone call.

The proposed system has short- and long-term benefits on the behavior of drivers. In the short-term, the proposed system along with a warning and/or ticketing protocol will make drivers think twice before reaching to their cell phones and focus on the task of driving. It is also possible to couple the warning system with the speed limiter subsystem of the vehicle. In the long run, not using cell phones while driving will become a social norm similar to the existing habit of wearing seat belts.

5. Conclusion

In this paper, we presented a new dataset for mobile cellphone usage detection. Also, we used WSN for the classification of the images in our presented dataset as well as the publicly available Kaggle State Farm dataset. We illustrated that the WSN outperforms the accuracy of well-known SOTA CNN models, but with a very small and simple architecture and far fewer parameters. We reached the accuracy of 91% for our domestic dataset and 99% for the State Farm publicly available dataset. We also showed that it can achieve even better results in image classification by training on grayscale images without image augmentations.

For future works, we will apply WSN for the localization of cellphone usage. Also, the same classification approach can be further improved by merging the datasets from the frontal view (like the dataset created by [24]) to the side view of the driver (Kaggle State Farm and dataset used by [14]). To compensate for the inability of the system in object detection in low light, architectures proposed by [29] and [28] are powerful solutions. For detection of other ways of using cellphones like hands-frees or speakers, an ensemble of vision-based techniques with methods like [37] and [54] may overcome the problem.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Y. Abouelnaga, *Real-time distracted driver posture classification*, in 32nd Conference on Neural Information Processing Systems (NIPS 2018), Workshop on Machine Learning for Intelligent Transportation Systems, 2017.

[2] S. Agrawal, A. M. Schuster, N. Britt, E. A. Mack, M. L. Tidwell, and S. R. Cotten, *Building on the past to help prepare the workforce for the future with automated vehicles: A systematic review of automated passenger vehicle deployment timelines*, Technology in Society, 72 (2023), p. 102186.

[3] M. Andreux, T. Angles, G. Exarchakis, R. Leonarduzzi, G. Rochette, L. Thiry, J. Zarka, S. Mallat, J. Andén, E. Belilovsky, J. Bruna, V. Lostanlen, M. Chaudhary, M. J. Hirn, E. Oyallon, S. Zhang, C. Cella, and M. Eickenberg, *Kymatio: Scattering transforms in python*, Journal of Machine Learning Research, 21 (2020), pp. 1–6.

[4] B. Baheti, S. Gajre, and S. Talbar, *Detection of distracted driver using convolutional neural network*, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1145–1151.

[5] A. B. Balbinot, M. A. Zaro, and M. I. Timm, *Funções psicológicas e cognitivas presentes no ato de dirigir e sua importância para os motoristas no trânsito*, Ciências & Cognição, 16 (2011), pp. 13–29.

[6] R. Berri, F. Osório, R. Parpinelli, and A. Silva, *A hybrid vision system for detecting use of mobile phones while driving*, in 2016 International Joint Conference on Neural Networks (IJCNN), 2016, pp. 4601–4610.

[7] S. Box, *Transportation institute releases findings on driver behavior and crash factors*. `https://vtechworks.lib.vt.edu/bitstream/handle/10919/59404/2006-237.html`, 2006. Last accessed 2022/11/23.

[8] D. J. Carragher, A. Towler, V. R. Mileva, D. White, and P. J. B. Hancock, *Masked face identification is improved by diagnostic feature training*, Cognitive Research: Principles and Implications, 7 (2022), pp. 1–12.

[9] F. Chollet, *Xception: Deep learning with depthwise separable convolutions*, in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 2016, Institute of Electrical and Electronics Engineers Inc, pp. 1800–1807.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[11] S. Ebrahimian, A. Nahvi, M. Tashakori, H. Salmanzadeh, O. Mohseni, and T. Leppänen, *Multi-level classification of driver drowsiness by simultaneous analysis of ecg and respiration signals using deep neural networks*, International Journal of Environmental Research and Public Health, 19 (2022).

[12] A. Eckert, A. Hohm, and S. Lueke, *An integrated adas solution for pedestrian collision avoidance*, in Proceedings of the 23rd International Conference on the Enhanced Safety of Vehicles, Seoul, Republic of Korea, 2013, pp. 13–298.

[13] Y. Elqattan, M. N. Moustafa, and M. H. El-Shafey, *System for detecting and reporting cell phone distracted drivers*, in 11th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 2019, IEEE, pp. 215–221.

[14] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, *Driver distraction identification with an ensemble of convolutional neural networks*, Journal of advanced transportation, 2019 (2019), p. 4125865.

[15] François Chollet, *Keras*. `https://github.com/fchollet/keras`, 2022. Last accessed 2022/12/19.

[16] Free Report: National Safety Council, *Understanding the distracted brain*, April 2012. WHY DRIVING WHILE USING HANDS-FREE CELL PHONES IS RISKY BEHAVIOR.

[17] C. Gröger, *There is no ai without data*, Commun. ACM, 64 (2021), pp. 98–108.

[18] D. Gruyer, S. Choi, C. Boussard, and B. d'Andréa Novel, *From virtual to reality, how to prototype, test and evaluate new adas: Application to automatic car parking*, in 2014 IEEE Intelligent Vehicles Symposium Proceedings, 2014, pp. 261–267.

[19] S. Hakak, T. R. Gadekallu, P. K. R. Maddikunta, S. P. Ramu, P. M, C. De Alwis, and M. Liyanage, *Autonomous vehicles in 5g and beyond: A survey*, Vehicular Communications, 39 (2023), p. 100551.

[20] A. He, G. Chen, W. Zheng, Z. Ni, Q. Zhang, and Z. Zhu, *Driver cell-phone use detection based on cornernet-lite network*, in IOP Conference Series: Earth and Environmental Science, vol. 632, IOP Publishing Ltd, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[22] V. Ilková and A. Ilka, *Legal aspects of autonomous vehicles — an overview*, in 2017 21st International Conference on Process Control (PC), 2017, pp. 428–433.

[23] I. Isaksson-Hellman and M. Lindman, *Real-world evaluation of driver assistance systems for vulnerable road users based on insurance crash data in sweden*, in 26th International Technical Conference on the Enhanced Safety of Vehicles (ESV): Enabling a Safer Tomorrow, NHTSA Washington, DC, 2019.

[24] P. Lassmann, M. S. Fischer, H.-J. Bieg, M. Jenke, F. Reichelt, G.-J. Tuezuen, and T. Maier, *Keeping the balance between overload and underload during partly automated driving: relevant secondary tasks*, in Automatisiertes Fahren 2019, T. Bertram, ed., Springer Fachmedien Wiesbaden, Wiesbaden, 2020, pp. 233–250.

[25] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, *Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016, pp. 46–53.

[26] S. Manimala and C. Kumar, *Anticipating hand and facial features of human body using golden ratio*, International Journal of Graphics & Image Processing (IJGIP), 4 (2014), p. 15.

[27] A. Montoya, D. Holman, SF-Data-Science, T. Smith, and W. Kan, *State farm distracted driver detection*. `https://kaggle.com/competitions/state-farm-distracted-driver-detection`, 2016. Last accessed 2022/12/17.

[28] I. Morawski, Y.-A. Chen, Y.-S. Lin, and W. H. Hsu, *Nod: Taking a closer look at detection under extreme low-light conditions with night object detection dataset*, 2021.

[29] R. Mukherjee, M. Bessa, P. Melo-Pinto, and A. Chalmers, *Object detection under challenging lighting conditions using high dynamic range imagery*, IEEE Access, 9 (2021), pp. 77771–77783.

[30] E. Oyallon, S. Mallat, and L. Sifre, *Generic deep networks with wavelet scattering*, in International Conference on Learning Representations (ICLR), 2014.

[31] M. Peissner, V. Doebler, and F. Metze, *Can voice interaction help reducing the level of distraction and prevent accidents?*, tech. rep., Fraunhofer IAO, 2011.

[32] PyTorch documentation, *Illustration of transforms – torchvision main documentation.* https://pytorch.org/vision/main/auto_examples/transforms/plot_transforms_illustrations.html#sphx-glr-auto-examples-transforms-plot-transforms-illustrations-py, 2022. Last accessed 2022/10/30.

[33] H. Qiu, D. Gong, Z. Li, W. Liu, and D. Tao, *End2end occluded face recognition by masking corrupted features*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (2022), pp. 6939–6952.

[34] P. Rajput, S. Nag, and S. Mittal, *Detecting usage of mobile phones using deep learning technique*, in Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good, GoodTechs '20, New York, NY, USA, 2020, Association for Computing Machinery, pp. 96–101.

[35] J. Redmon and A. Farhadi, *Yolov3: An incremental improvement*, 2018.

[36] M. Regan, T. Victor, J. Lee, and K. Young, *Driver distraction injury prevention countermeasures—part 3: vehicle, technology, and road design*, in Driver Distraction: Theory, Effects, and Mitigation, CRC Press, 2009, pp. 579–601.

[37] J. M. Rodríguez-Ascariz, L. Boquete, J. Cantos, and S. Ortega, *Automatic system for detecting driver use of mobile phones*, Transportation Research Part C: Emerging Technologies, 19 (2011), pp. 673–681.

[38] T. U. Saeed, *Road infrastructure readiness for autonomous vehicles*, PhD thesis, Purdue University, 2019.

[39] V. Saini and R. Saini, *Driver drowsiness detection system and techniques: A review*, International Journal of Computer Science and Information Technologies, 5 (2014), pp. 4245–4249.

[40] A. Salvador, *A culpa foi do celular*, Revista Veja, 21 (2011), p. 2011.

[41] K. Sandeep, P. Ravikumar, and S. Ranjith, *Novel drunken driving detection and prevention models using internet of things*, in 2017 International Conference on Recent Trends in Electrical, Electronics and Computing Technologies (ICRTEECT), 2017, pp. 145–149.

[42] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor, *Driver cell phone usage detection on strategic highway research program (shrp2) face view videos*, in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 35–43.

[43] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, *Real-time rotation-invariant face detection with progressive calibration networks*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2295–2303.

[44] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, Computational and Biological Learning Society, 2015, pp. 1–14.

[45] A. Stevens and R. Minton, *In-vehicle distraction and fatal accidents in england and wales*, Accident Analysis & Prevention, 33 (2001), pp. 539–545.

[46] H. B. Sundfør, F. Sagberg, and A. Høye, *Inattention and distraction in fatal road crashes–results from in-depth crash investigations in norway*, Accident Analysis & Prevention, 125 (2019), pp. 152–157.

[47] P. Taunk, G. Jayasri, J. Priya, and N. S. Kumar, *Face detection using viola jones with haar cascade*, Test Engineering and Management, 83 (2020), pp. 19146–19149.

[48] R. Torres, O. Ohashi, E. Carvalho, and G. Pessin, *A deep learning approach to detect distracted drivers using a mobile phone*, in Artificial Neural Networks and Machine Learning – ICANN 2017, Belém, PA, Brazil, 2017, Springer Science, Springer International Publishing, pp. 72–79.

[49] A. Validi, T. Ludwig, A. Hussein, and C. Olaverri-Monreal, *Examining the impact on road safety of different penetration rates of vehicle-to-vehicle communication and adaptive cruise control*, IEEE Intelligent Transportation Systems Magazine, 10 (2018), pp. 24–34.

[50] B. Wagner, F. Taffner, S. Karaca, and L. Karge, *Vision based detection of driver cell phone usage and food consumption*, IEEE Transactions on Intelligent Transportation Systems, 23 (2022), pp. 4257–4266.

[51] Q. Xiong, J. Lin, W. Yue, S. Liu, Y. Liu, and C. Ding, *A deep learning approach to driver distraction detection of using mobile phone*, in 2019 IEEE Vehicle Power and Propulsion Conference (VPPC), 2019, pp. 1–5.

[52] B. Xu and R. P. Loce, *A machine learning approach for detecting cell phone usage*, in Video Surveillance and Transportation Imaging Applications 2015, R. P. Loce and E. Saber, eds., vol. 9407, International Society for Optics and Photonics, SPIE, 2015, p. 94070A.

[53] B. Xu, P. Paul, Y. Artan, and F. Perronnin, *A machine learning approach to vehicle occupancy detection*, in 17th IEEE International Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 2014, Institute of Electrical and Electronics Engineers Inc, pp. 1232–1237.

[54] J. Yang, S. Sidhom, G. Chandrasekaran, T. Vu, H. Liu, N. Cecan, Y. Chen, M. Gruteser, and R. P. Martin, *Detecting driver phone use leveraging car speakers*, in Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, MobiCom '11, New York, NY, USA, 2011, Association for Computing Machinery, pp. 97–108.

[55] H. Yasar, *Detection of driver's mobile phone usage*, in 2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2017, pp. 1–4.

[56] Y. Zhang, C. Fu, and S. Cheng, *Exploring driver injury severity at intersection: An ordered probit analysis*, Advances in Mechanical Engineering, 7 (2015), p. 567124.

[57] R. Zhou, Y. Zhang, and Y. Shi, *Driver's distracted behavior: The contribution of compensatory beliefs increases with higher perceived risk*, International Journal of Industrial Ergonomics, 80 (2020), p. 103009.

Please cite this article using:

Ali Besharati, Ali Nahvi, Serajeddin Ebrahimian, Driver cellphone usage detection using wavelet scattering and convolutional neural networks, AUT J. Math. Comput., 6(3) (2025) 257-268
https://doi.org/10.22060/AJMC.2023.22580.1177