

## DATA AUGMENTATION FOR A LEARNING-BASED VEHICLE MAKE-MODEL AND LICENSE PLATE MATCHING SYSTEM


**Burak AĞGÜL<sup>1\*</sup>, Gökhan ERDEMİR<sup>2</sup>**


*The most important requirement for deep learning algorithms to run with a low error ratio is the realization of the training process with a sufficient amount of data. Using synthetic data is one of the most common approaches when the data set is not enough for training. Synthetic data production must be based on a real dataset to improve the prediction and classification abilities of the deep learning algorithms. The enrichment of the existing dataset using different techniques such as modified copies of existing data is called data augmentation. It can sometimes be difficult to generate enough datasets according to the type of problem, especially in image classification. In such cases, a dataset can be generated by duplicating and/or modifying existing pictures of the objects. In this study, data augmentation for a learning-based vehicle make-model and license plate matching system has been performed and a new vehicle image dataset has been generated. The proposed approach which has been used in creating the dataset is presented in detail. The generated new vehicle image dataset is available to developers as open-source.*

*Key words: data augmentation, deep learning, dataset, open-source, synthetic data generation*

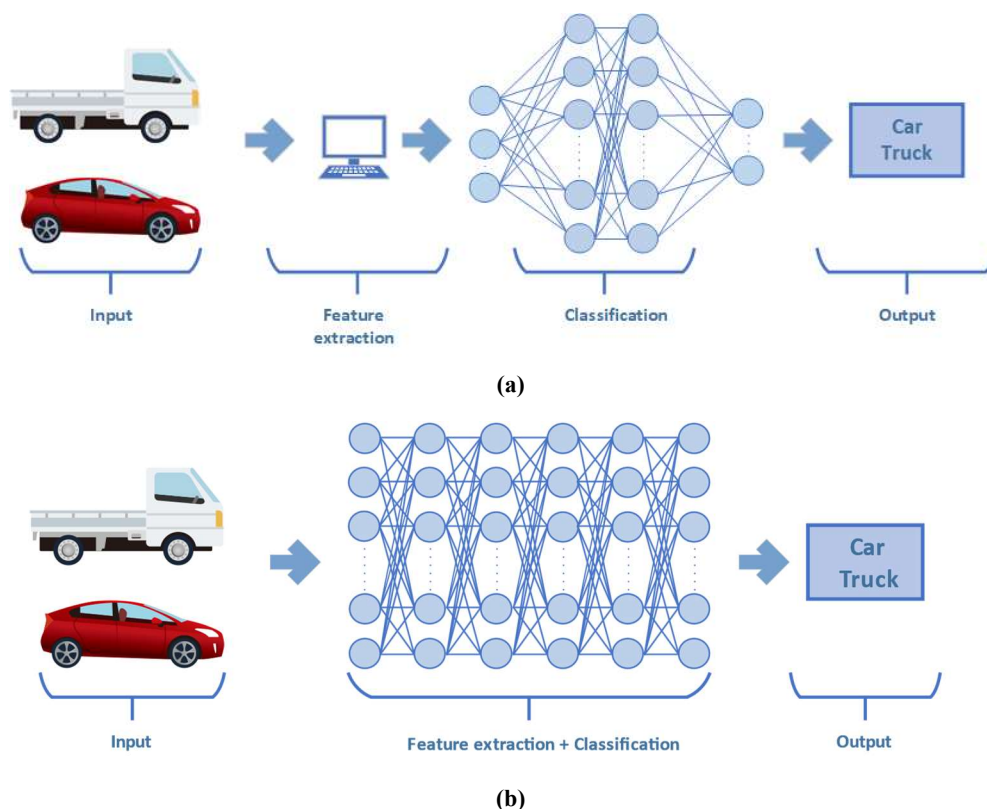
### 1. Introduction

In this study, all stages of the data set generation which can be used for the vehicle make-model and license plate matching system are presented detail. In general, learning-based software systems are used to monitor and identify vehicles in traffic instantly. One of the most essential requirements for learning-based prediction and identification systems to run with a low error ratio is the realization of the training process with a sufficient amount of data [1–3]. Enough amount of data used for learning and testing increases the accuracy of the training stage [4–6]. Hence, learning-based algorithms may be capable of detecting small details and differences [7–9]. Moreover, the qualities and content of the datasets differ according to the aim of the studies. However, the main problem in this approach is to take and store different vehicle images to generate datasets. According to our literature review on vehicle make-model datasets, it has been observed that the existing datasets do not have enough vehicle pictures not only front-view but also rear-view for deep learning[10–13]. It has been observed that a few studies were focused on to develop datasets for vehicle images.

<sup>1</sup> Department of Computer Science and Engineering, Istanbul Sabahattin Zaim University, Istanbul, Turkey, (aggul.burak@std.izu.edu.tr).  <https://orcid.org/0000-0002-9183-1568>

<sup>2</sup> Department of Electrical and Electronics Engineering, Istanbul Sabahattin Zaim University, Istanbul, Turkey, (gokhan.erdemir@izu.edu.tr)  <https://orcid.org/0000-0003-4095-6333>

In recent years, one of the most important innovations in software development is deep learning which is a machine learning method that expresses deep neural networks[14–17]. It is an intelligent artificial neural network (ANN) system that allows us to train the created models to predict outputs using artificial neural networks from a given data set. But the main difference between ANN and deep learning is that a lot of data is needed for the application of deep learning. In the other words, deep learning is a more advanced form of machine learning method which needs a lot of data to find and to model correlations between inputs and outputs in the training stage. The other important point is that it creates new features by learning from the data itself. The difference between machine learning and deep learning for vehicle identification systems is shown in Figure 1. as block diagrams. In machine learning, determined features must be defined or created by the user before the training stage. On the other hand, in deep learning, features are defined or created by itself using ANN. The deep learning system identifies and tags extracted features and then starts to produce outputs using ANN, as well. The size of the data set is one of the important points when designing such systems. While designing a learning-based vehicle make-model and license plate matching system using deep learning methods, the system must be based on digital images, so the number of data should be expressed in thousands or more, not in hundreds.



**Figure 1. (a) Machine learning and (b) deep learning block diagrams.**

In this study, two different vehicle brands, Honda and Ford, which referred to herein as H and F are used. If data will be classified as two different objects (or vehicles for this study), it is necessary to create datasets that have an almost equivalent number of data. For example, let's assume that the H and the F brands have 2000 and 5000 data in their own datasets, respectively. It is not possible to make training and tests with these kinds of datasets for deep learning algorithms, accurately. Hence, approximately the same size datasets should be used for the learning stage. This approach increases the performance of the algorithms and decreases the prediction error ratio.

To increase the size of the dataset does not mean that the success rate of the model will increase linearly. Model accuracy can be stable, or its performance can be enhanced with small ratios such as one-thousandth, etc. In these cases, when the performance ratio of the model that continues with enhancements like one-thousandth is observed, at this point, the number of data in datasets should be fixed after an obvious breakpoint is determined. On the other hand, the contrary situation is possible. In this circumstance, the logical way is to use synthetic and/or augmented data for training and testing of the model. Thus, the performance of training and testing of the model can be approached to the desired ratio if it is possible as well by using synthetic and/or augmented data. As a result, it is possible to generate datasets that include thousands of data with different quantities and contents.

In this study, data augmentation for a learning-based vehicle make-model and license plate matching system has been performed and vehicle image datasets that belong to the F and the H brands have been generated. The generated datasets with augmented data are available to developers as open-source. Developers can download them from the following website [18]. The paper is organized as the following outlines. The preparation of datasets, data augmentation approaches, and properties of datasets are presented, in section II. In Section III, new datasets are presented. Finally, the conclusion is presented in the last section.

## 2. Materials and methods

### 2.1. Dataset Preparation

In this study, two datasets that belong to the Focus(F) model of the Ford(F) between the 2012-2014 model year and the Civic(C) model of the Honda(H) brand with the 2016-2019 model year have been generated. Each dataset contains not only the front view of the vehicle but also the rear view of the vehicle. The block diagram of the data preparation steps is presented in Figure 2. As a first step, the low-resolution images obtained from the vehicle sales website were recorded automatically while generating the datasets. Each image has been tagged according to the year-brand-model pattern as a second step.



Figure 2. Block diagram of data preparation steps.

For example, the tag *2012\_2014\_F\_F\_F* was used for the front view of the F model of the F brand. The tag *2012\_2014\_F\_F\_R* was used for the rear view of the same vehicle. the tag *2016\_2019\_H\_C\_F* was used for the front view of the C model of the H brand. The tag *2016\_2019\_H\_C\_R* was used for the rear view of the same vehicle. Thus, 2 different classes were generated for each vehicle model. Each image which has been download from [19] and [20] has 600x450 pixel dimensions.

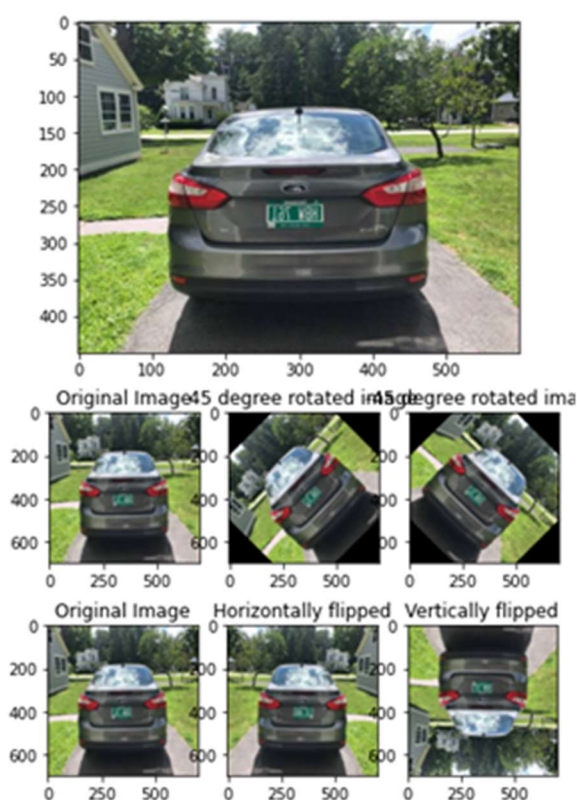
### 2.2. Image augmentation process

In the literature, data augmentation which produces synthetic data from one or more real data is the morphological process[1,4,9,21]. In this progress, it is the acquisition of new images by applying processes such as rotation, horizontal and vertical inversion, translation, scaling, shifting, convert to

grayscale, scrolling, Gauss noise (adding pixels), and etc. to existing[22]. The small datasets can be transformed into a larger training set with the new synthetic data created in this way. And also, approximately 40 random new images can be obtained from an image. Thus, overlapping in the training stage can be prevented due to the images which were produced using this technique[14,17]. In this study, 10000 images were generated by using rotation, horizontal inversion, and vertical inversion techniques. These morphological processes can be explained as follows:

- **Rotation:** It is an important point to note with the rotation process is that the image dimensions may not be preserved after rotation. If an image has square dimensions, rotating it 90 degrees keeps the image size the same. If it has rectangular dimensions, rotating it 180 degrees keep the size the same.
- **Horizontal inversion (flipped):** It is the complete reversal of the image on the horizontal axis. For example, it is the process of converting the vehicle image taken from the left side to the right side as a mirror reflection.
- **Vertical inversion (flipped):** It is the complete reversal of the image on the vertical axis.

Generated images by rotation, vertical inversion, and horizontal inversion processes from the original image are presented in Figure 3.



**Figure 3.** New images were generated by using rotation, vertical inversion and horizontal inversion processes from the original image.

- **Translation:** It involves moving the image in the X or Y direction (or both).
- **Shifting:** An image is shifted to the left or right by a selected pixel ratio on the x-axis. The overflow part of the image is added again from the other side of the image.
- **Convert to grayscale:** All colored pixels convert to grayscale.

Generated images by shifting and convert to grayscale processes from the original image are presented in Figure 4.

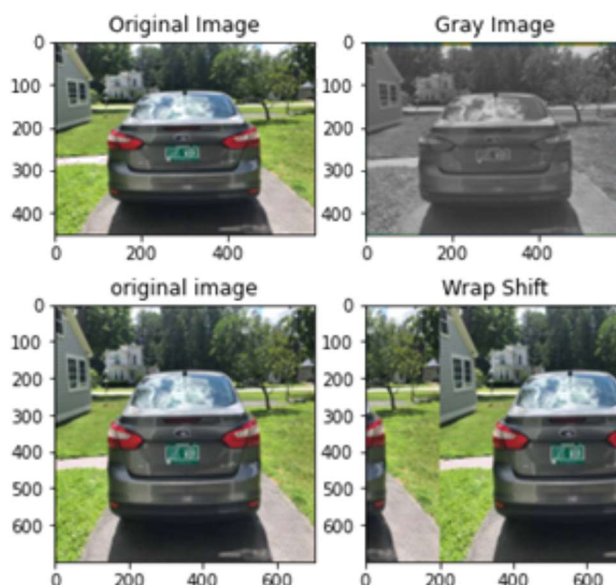


Figure 4. New images were generated by using rotation, vertical inversion and horizontal inversion processes from the original image.

- **Gauss noise (adding pixels):** Gaussian noise with a zero average has data points at essentially all frequencies and effectively distorts high-frequency characteristics. Adding enough amount of noise can increase learning ability.
- **Blurring:** A filter used for blurring is a low pass filter. It allows to pass the low-frequency and cut high-frequency. A blurred image has not got sharp edges. Thus, this kind of filter is preventing the model from over-fitting.

Generated images by adding noise (Gauss noise) and blurring processes from the original image are presented in Figure 5.

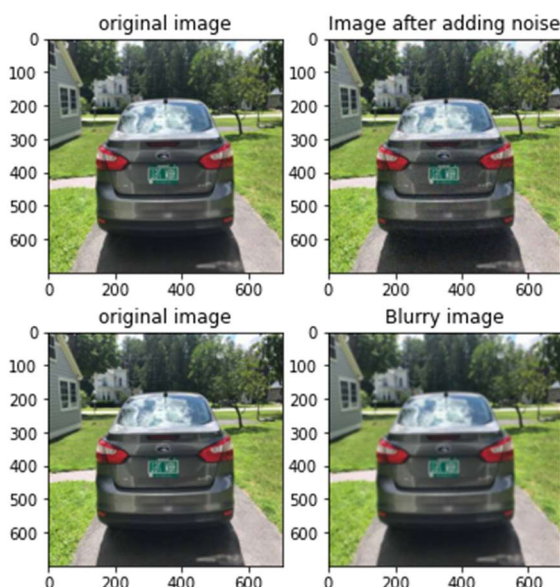


Figure 5. New images were generated by adding noise (Gauss noise) and blurring.

- **Cropping:** It can be used to get the sample a random part from the original image. Then, the cropping part of the image is scaled to the original size of the image. This method is commonly known as random cropping.

- **Scaling:** The image can be scaled outward or inward by this technique. The dimension of the generated images is larger than the original image, possibly. If this situation occurs, it is necessary to equal the new image size to the original image.

### 2.3. Clearing similar images

It is possible to produce similar synthetic images due to random generation. And also, sometimes, some images can be exactly the same. To have similar images in the dataset may cause overlapping. Therefore, it is necessary to clear similar images from datasets. In this study, similar images (size, scale, etc.) removed from datasets by using a recursive clearing algorithm which compares all generated images recursively.

### 2.4. Tagging image names

Each augmented image was tagged with names to match the folder name in the entire datasets. Samples of augmented and tagged images are presented in Figure 6. As seen in Table 1, 10000 training, 1000 test, and 100 verification images were created in order to perform training, testing, and validation. Samples from the generated dataset are shown in Figure 7.

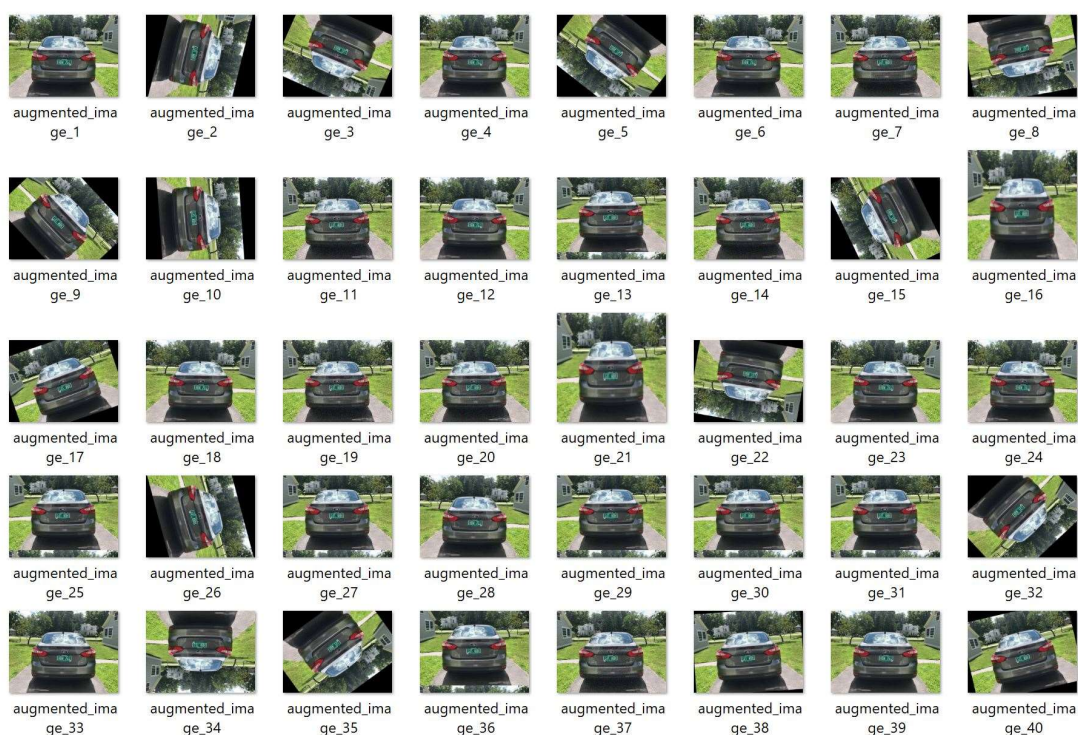


Figure 6. Samples of augmented images.

Table 1. Properties of the generated datasets.

Tag (Year/Brand/Model/View side)	Training	Test	Validation
2012_2014_F_F_F	2500	250	25
2012_2014_F_F_R	2500	250	25
2016_2019_H_C_F	2500	250	25
2016_2019_H_C_R	2500	250	25

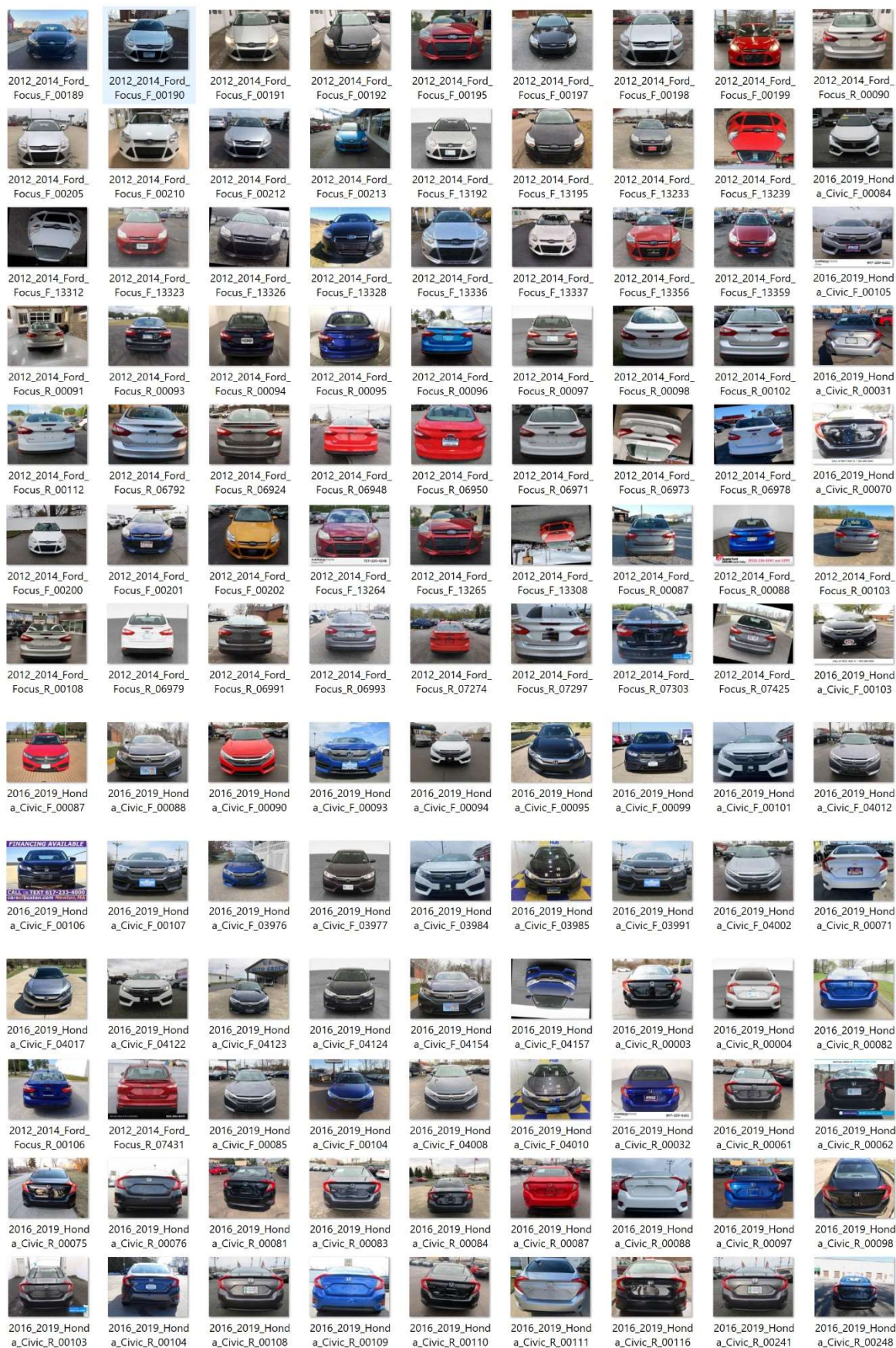


Figure 7. Samples from the generated dataset.

### 3. Conclusion

In this study, the steps of creating a dataset for a learning-based vehicle make-model and license plate matching system are proposed. For the generated dataset content, it was created as 4 classes based on one model of each of the two-vehicle brands. The dataset was generated as that can be seen in the brand model of the vehicle, as well as model year, front-view, and rear-view. The proposed approach which has been used in creating the dataset is presented in detail. The generated new vehicle image dataset is available to developers as open-source.

### References

- [1] F. Cen, X. Zhao, W. Li, G. Wang, Deep Feature Augmentation for Occluded Image Classification, *Pattern Recognit.* 111 (2020) 107737. <https://doi.org/10.1016/j.patcog.2020.107737>.
- [2] H.C. Shin, K. Il Lee, C.E. Lee, Data augmentation method of object detection for deep learning in maritime image, *Proc. - 2020 IEEE Int. Conf. Big Data Smart Comput. BigComp 2020.* (2020) 463–466. <https://doi.org/10.1109/BigComp48618.2020.00-25>.
- [3] H. Zheng, H. Shang, Z. Sun, X. Fu, J. Yao, J. Huang, Supervised Augmentation: Leverage Strong Annotation for Limited Data, *Proc. - Int. Symp. Biomed. Imaging. 2020-April* (2020) 1134–1138. <https://doi.org/10.1109/ISBI45749.2020.9098607>.
- [4] D. Zhao, G. Yu, P. Xu, M. Luo, Equivalence between dropout and data augmentation: A mathematical check, *Neural Networks.* 115 (2019) 82–89. <https://doi.org/10.1016/j.neunet.2019.03.013>.
- [5] A. Sakai, Y. Minoda, K. Morikawa, Data augmentation methods for machine-learning-based classification of bio-signals, *BMEiCON 2017 - 10th Biomed. Eng. Int. Conf. 2017-January* (2017) 1–4. <https://doi.org/10.1109/BMEiCON.2017.8229109>.
- [6] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, *2018 Int. Interdiscip. PhD Work. IIPHDW 2018.* (2018) 117–122. <https://doi.org/10.1109/IIPHDW.2018.8388338>.
- [7] J. Nalepa, G. Mrukwa, S. Piechaczek, P.R. Lorenzo, M. Marcinkiewicz, B. Bobek-billewicz, P. Wawrzyniak, P. Ulrych, J. Szymanek, M. Cwiek, W. Dudzik, M. Kawulok, M.P. Hayball, DATA AUGMENTATION VIA IMAGE REGISTRATION Future Processing , Gliwice , Poland Institute of Informatics , Silesian University of Technology , Gliwice , Poland Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology , Gliwice , Poland *Feedba*, (2019) 4250–4254.
- [8] H. Li, J. Rao, L. Zhou, J. Zhang, Valid data augmentation by patch alpha matting, *2019 IEEE 4th Int. Conf. Signal Image Process. ICSIP 2019.* (2019) 361–366. <https://doi.org/10.1109/SIPROCESS.2019.8868572>.
- [9] J. Nalepa, M. Myller, M. Kawulok, Training- And Test-Time Data Augmentation for Hyperspectral Image Segmentation, *IEEE Geosci. Remote Sens. Lett.* 17 (2020) 292–296. <https://doi.org/10.1109/LGRS.2019.2921011>.
- [10] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, F. Mujica, A. Coates, A.Y. Ng, An Empirical Evaluation of Deep Learning on Highway Driving, (2015) 1–7. <http://arxiv.org/abs/1504.01716>.
- [11] A.T. Sasongko, G. Jati, M.I. Fanany, W. Jatmiko, Dataset of vehicle images for Indonesia toll road tariff classification, *Data Br.* 32 (2020) 106061. <https://doi.org/10.1016/j.dib.2020.106061>.
- [12] E. Osaba, Benchmark dataset for the Asymmetric and Clustered Vehicle Routing Problem with



- Simultaneous Pickup and Deliveries, Variable Costs and Forbidden Paths, *Data Br.* 29 (2020) 105142. <https://doi.org/10.1016/j.dib.2020.105142>.
- [13] W. Yang, Z. Li, C. Wang, J. Li, A multi-task Faster R-CNN method for 3D vehicle detection based on a single image, *Appl. Soft Comput. J.* 95 (2020) 106533. <https://doi.org/10.1016/j.asoc.2020.106533>.
- [14] J. Arun Pandian, G. Geetharamani, B. Annette, Data Augmentation on Plant Leaf Disease Image Dataset Using Image Manipulation and Deep Learning Techniques, *Proc. 2019 IEEE 9th Int. Conf. Adv. Comput. IACC 2019.* (2019) 199–204. <https://doi.org/10.1109/IACC48062.2019.8971580>.
- [15] N. Varela, C.G. Zoe, R. Ternera-Muñoz Yesith, F. Esmeral-Romero Ernesto, N.A.L. Zelaya, Method for classifying images in databases through deep convolutional networks, *Procedia Comput. Sci.* 175 (2020) 135–140. <https://doi.org/10.1016/j.procs.2020.07.022>.
- [16] Y. Bengio, Learning deep architectures for AI, *Found. Trends Mach. Learn.* 2 (2009) 1–27. <https://doi.org/10.1561/22000000006>.
- [17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 15 (2018) 7642–7651. <https://doi.org/10.1109/CVPR.2018.00797>.
- [18] B. Ağgül, GitHub, (2020). <https://github.com/burakaggul/Vehicle-brand-model-recognition-with-deep-learning-using-keras>.
- [19] Craigslist.org, (2020). <https://cfl.craigslist.org/>.
- [20] Autoscout24, (2020). <https://www.autoscout24.com.tr/>.
- [21] J. Shijie, W. Ping, J. Peiyi, H. Siping, Research on data augmentation for image classification based on convolution neural networks, *Proc. - 2017 Chinese Autom. Congr. CAC 2017.* 2017-January (2017) 4165–4170. <https://doi.org/10.1109/CAC.2017.8243510>.
- [22] T.D. Pham, Geostatistical Simulation of Medical Images for Data Augmentation in Deep Learning, *IEEE Access.* 7 (2019) 68752–68763. <https://doi.org/10.1109/ACCESS.2019.2919678>.