Building a Driving Behaviour Dataset

Ion Cojocaru
University of Craiova,
13 A.I.Cuza Street, 200585
Craiova, Romania
kojocaru.ivan@gmail.com

Paul-Stefan Popescu
University of Craiova,
13 A.I.Cuza Street, 200585
Craiova, Romania
stefan.popescu@edu.ucv.ro

ABSTRACT
Traffic has become more complex, and ride-sharing applications are used more often. Once with them, the problem of detecting and understanding the driving behaviour became more relevant. The problem is that now we tried to build a system for analysing the driving behaviour, we were unable to find a dataset that satisfies two important requirements: to be labelled (not just collected data) and ready to be used in a model designed to be integrated into the smartphones' applications. In this paper, we present a dataset collected using an Android smartphone that uses only the smartphone's sensors data. The dataset is labelled in three classes: slow, normal and aggressive and is described along with experiments designed to provide an insight into the data quality.

Author Keywords
Machine learning; dataset; driving dataset, driving behaviour.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors; Design; Measurement.

DOI: 10.37789/rochi.2022.1.1.17

INTRODUCTION

Driving behaviour is a complex concept describing how the driver operates the vehicle in the context of the driving scene and surrounding environment. It's one of the essential aspects in the design, development, and application of Advanced Driving Assistance Systems [1] and Intelligent Transportation Systems [2], which can be affected by many factors. Correct analysis and measuring of driving styles could significantly improve a driver's comfort and experience.

Nowadays, almost everyone has a mobile phone. It is an excellent communication tool and, most importantly, affordable for nearly everyone. Considering this, we could say that a smartphone is the best tool for our task. A modern smartphone has many sensors, starting with a classical accelerometer and ending with a recently introduced soli sensor.

For our purpose, we'll need sensors that will keep track of a car's movement, like an accelerometer and gyroscope, and we choose to use only these two sensors in order to have a generic approach which fits most the situation and also can be used with as many as possible devices.

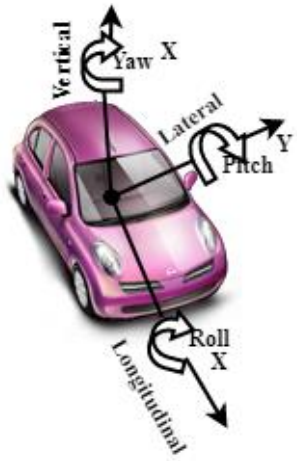

Figure 1. Car example

Figure 1 presents a car along with the three axes in which the movement is possible. As we can observe, there are three-axis that we are interested in keeping track of longitudinal, lateral and vertical events. An accelerometer sensor for monitoring sudden movements and a gyroscope for orientation will do the trick. These sensors provide the exact data we are looking for on the X, Y and Z axes.

We consider logging the data from these sensors because they offer data relevant for vehicle movement, and we can compute how strong the forces applied to the car. The motivation for using the data collected from these sensors rather than GPS, which can provide speed, is that an average or slow driver may drive at high speed where the road allows but will not make sudden moves. In our approach, we try to find the sudden movements made by the driver and, based on this data, detect the driving style.

The proposed dataset[1] was collected using a Samsung Galaxy S10 smartphone and a Dacia Sandero 1.4 MPI,

[1] https://www.kaggle.com/datasets/outofskills/driving-behavior

which had 75 horsepower. The code is available on Github[2] and can be used for further collecting data and improving the dataset. Regarding the car, it was chosen to be a regular one with a mainstream engine that ride-sharing drivers usually select. A more powerful car would differentiate between driving styles better and make the machine learning task less complicated. Still, if the model is used on a slower car, the results may not be that reliable. The dataset is composed of two files, one for training and one for testing, each having the same structure, so they can also be merged for training a better model. We choose this approach because we need to see how good the model is when driving on a different route, so we will be able to evaluate how generic this approach is.

RELATED WORK

Regarding driving behaviour datasets, there are already some of them on Kaggle or Mendeley. Still, despite the collected data, they all have a big problem: the data is not labelled. The lack of labelling makes the users and data analysis guess which driving behaviour was adopted, and this implies another big problem, how can we scale the data to the car's capabilities. In this direction, there is a new paper [3] which describes a similar approach with data collected from gyroscope and accelerometer, but the data [3] is unlabelled and difficult to use for machine learning or deep learning model training. The data is organised into day-by-day folders, each with seven subfolders. The authors state that they are confident that the suggested dataset will be beneficial in the training, testing, and validation of a machine learning model for driver behaviour classification or reorganisation.

Another dataset[4] that was analysed before creating our own was the one described in [4], and the authors state that Machine learning techniques can enhance research, but these rely on large amounts of data which are difficult and very costly to obtain through Naturalistic Driving Studies (NDSs) [5], resulting in limited accessibility to the general research community. They also observe that the proliferation of smartphones has provided a cheap and easy-to-deploy platform for driver behaviour sense. Still, existing applications do not provide open access to their data. For these reasons, they wrote the paper that presents the UAH-DriveSet, a public dataset that allows deep driving analysis by providing a large amount of data captured by their driving monitoring app, DriveSafe. Their application is run by six different drivers and vehicles, performing 3 different behaviours (normal, drowsy and aggressive) on two types of roads (motorway and secondary road), resulting in more than 500 minutes of naturalistic driving with its associated raw data and processed semantic information, together with the video recordings of the trips. The dataset is good and comprehensive but also lacks labelling and is hard to be used in a different system, other than their own.

Honda Research Institute Driving Dataset (HDD)[5] is rather a database, not a real dataset which can be used for training an algorithm and is described in paper [6] but it's usage in an external application is quite difficult. The dataset includes 104 hours of real human driving in the San Francisco Bay Area collected using an instrumented vehicle equipped with different sensors. They provide a detailed analysis of HDD with a comparison to other driving datasets. A novel annotation methodology is introduced to enable research on driver behaviour understanding from untrimmed data sequences. As the first step, baseline algorithms for driver behaviour detection are trained and tested to demonstrate the feasibility of the proposed task.

To better understand the data from this domain we also explored the papers that analyse datasets that explore and classify the driver behaviour. One of these papers is [7] which presents A Recognition Method of Aggressive Driving Behaviour Based on Ensemble Learning. The authors found that there are some disadvantages, such as high miss rate and low accuracy, in the previous data-driven recognition methods of Aggressive Driving Behaviour, which are caused by the problems such as the improper processing of the dataset with imbalanced class distribution and one single classifier utilised. Aiming to deal with these disadvantages, an ensemble learning-based recognition method of Aggressive Driving Behaviour is proposed in their paper. First, the majority class in the dataset is grouped employing the self-organising map and then combined with the minority class to construct multiple class balance datasets. After that, three deep learning methods, including convolutional neural networks, long short-term memory, and gated recurrent units, are employed to build the base classifiers for the class balance datasets. In the end, the ensemble classifiers are combined with the base classifiers and then trained and verified using a multi-source naturalistic driving dataset acquired by the integrated experiment vehicle. The results suggest that in terms of the recognition of Aggressive Driving Behaviour, the ensemble learning method proposed in this research achieves better performance in all the metrics used than the aforementioned typical deep learning methods. Among the ensemble classifiers, the one based on the LSTM and the Product Rule has the optimal performance, and the other one based on the LSTM and the Sum Rule has the suboptimal performance

Earlier papers [8] use GPS as the data source and aim to create Driver behaviour profiles which are introduced in their paper as an approach for evaluating driver behaviour

[2] https://github.com/OutofSkills/AndroidDriverApp

[3] https://data.mendeley.com/datasets/9vr83n7z5j/2

[4] http://www.robesafe.uah.es/personal/eduardo.romera/uah-driveset/#dataset

[5] https://paperswithcode.com/dataset/hdd

as a function of the risk of a casualty crash. Their paper details the development of these Driver Behaviour Profiles and demonstrates their use as an input into modelling the factors that influence driver behaviour. The results show that even having controlled for the influence of the road environment, these factors remain the strongest predictors of driver behaviour, suggesting different spatiotemporal environments elicit various psychological responses in drivers.

A more advanced system (LIBRE: The Multiple 3D LiDAR Dataset) is presented in the paper [9]. As the authors state, the dataset is a first-of-its-kind dataset featuring ten different LiDAR sensors, covering a range of manufacturers, models, and laser configurations. Because of the advanced sensors used, the data captured independently from each sensor includes three different environments and configurations: static targets, where objects were placed at known distances and measured from a fixed position within a controlled environment; adverse weather, where static obstacles were calculated from a moving vehicle, captured in a weather chamber where LiDARs were exposed to different conditions (fog, rain, intense light); and finally, heavy traffic, where dynamic objects were captured from a vehicle driven on public urban roads, multiple times at different times of the day, and including supporting sensors such as cameras, infrared imaging, and odometry devices.

DATASET DESCRIPTION

Labelling

Collecting data to train a machine learning or deep learning model can be tricky as the results directly depend on the collected data's quality. By quality, it is assumed that the collected data is relevant to the assigned label.

In our case, we are assuming that a driving behaviour can be in one of the following classes:

- Aggressive – sudden left or right turns, acceleration and brake.
- Normal – average driving events.
- Slow – maintaining a lower-than-average speed

Starting from this assumption, we've driven three times, one aggressive, one normal and one slow ride on the same portion of the road.

Structure and routes

The training file is composed of 3644 recordings; 1331 are classified as slow, 1113 as aggressive, and 1200 as normal. Each set of instances is collected on the same road section. The number of instances varies between classes because we selected two readings per second, and driving aggressive takes less time than driving normal and slow. We have almost the same proportion of instances for each of the two datasets.

Figure 2. The route used for the training dataset

Figure 2 presents the route used for collecting the training dataset. This route was chosen carefully because we had to simulate city driving but driving aggressive or excessively slowly can be challenging during heavy traffic conditions, so we had to choose a route that would allow us to collect as relevant data as possible. The data collection round for each class was performed back and forth, having the same starting point.

The test dataset is composed of a slightly lower number of instances: 3084 divided into 1273 for slow, 997 for normal and 814 for aggressive, which reflects the proportions from the training dataset but with lower values. The route had a smaller distance and took less time for logging, influencing the number of instances collected.

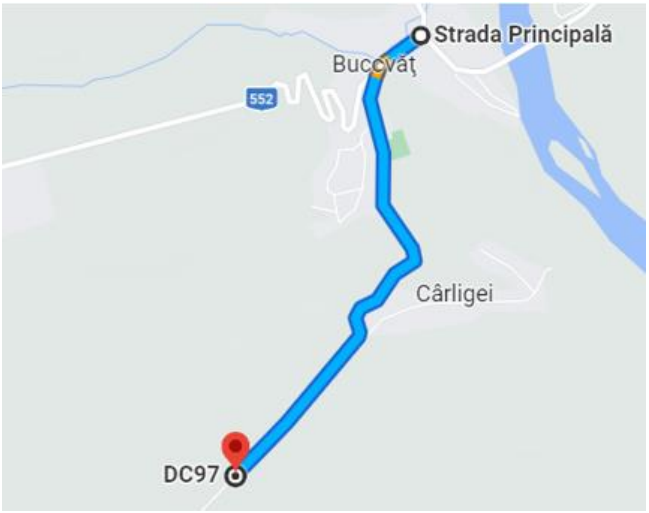

Figure 3. The route used for the test dataset

Figure 3 presents the route chosen for collecting the test dataset. The starting point is the same chosen for the training dataset, but the direction was different, and the road had similar characteristics. Even the daytime and

climate were similar to the ones from the training dataset so that the results will be as relevant as possible.

Data Analysis

For data analysis purposes, we will present how data is distributed and how the sensor's values differ between the driving behaviours collected. All the results and the charts added in this section can be reproduced using a *dense-nn*[6] notebook from Kaggle. The axis presented in the figures is the ones mentioned in Figure 1.

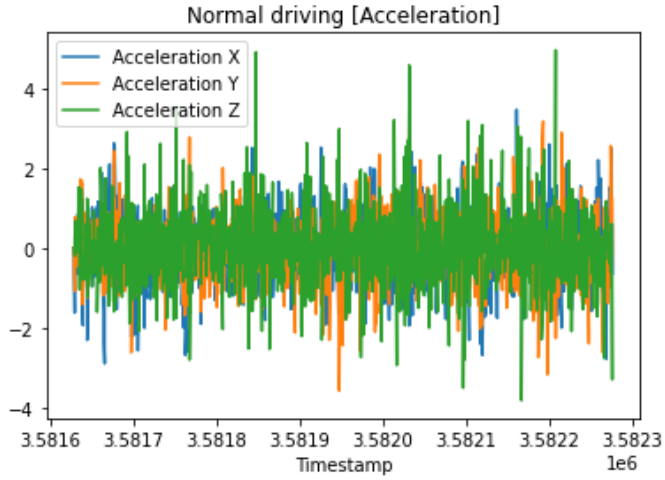

Figure 4. Accelerometer values for normal driving

Figure 4 presents the sensor values collected using an accelerometer. We need to consider that on the Z axis, we will always have plenty of noise because of road imperfections. The X axis is the most relevant for acceleration and breaking, and we can see how it blends with lateral forces, which are represented by the Y axes.

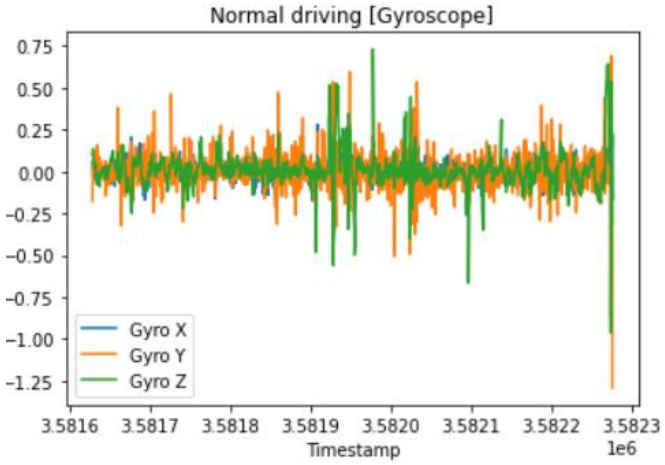

Figure 5. Gyroscope values for normal driving

Regarding the sudden turns and the gyroscope is the most relevant sensor. We will get rotation around Z and Y axes because a rotation around the X axes is almost impossible if the car is not rolling with the wheels upside down.

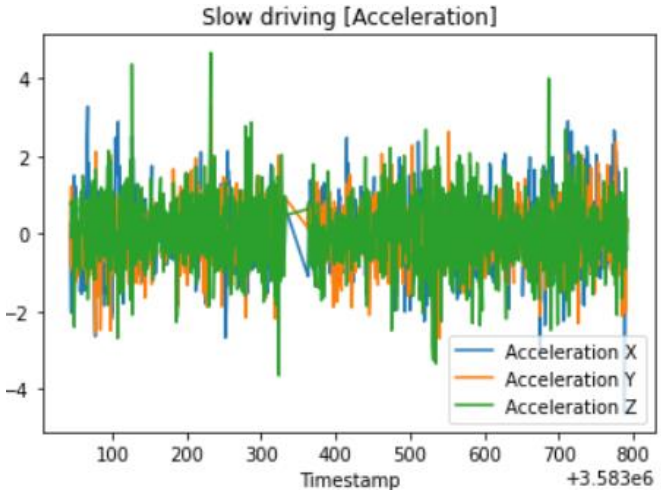

Figure 6. Accelerometer values for slow driving

Figure 6 represents the data collected for slow driving. Even though it has several similarities with Figure 4, we can see that acceleration on the X axis is less dominant because the brakes and acceleration were used more gently.

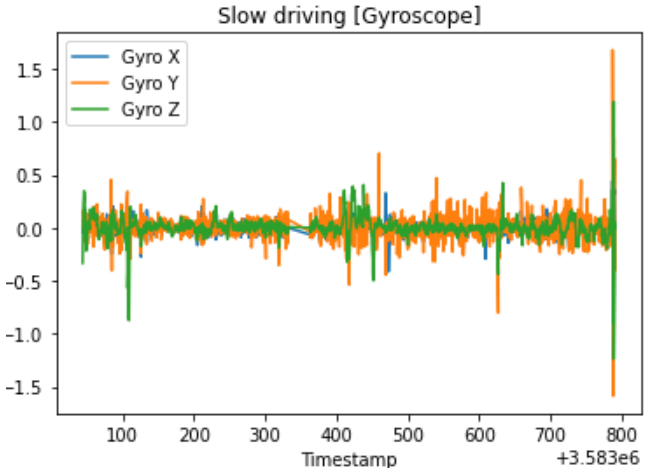

Figure 7. Gyroscope values for slow driving

Figure 7 presents the gyroscope values collected for slow driving and here we can see more differences between this figure and Figure 5 rather than the differences between Figure 4 and 6. We can observe here that despite the noise from the beginning and the end we had a smaller amplitude.

[6] Charts: https://www.kaggle.com/code/outofskills/dense-nn

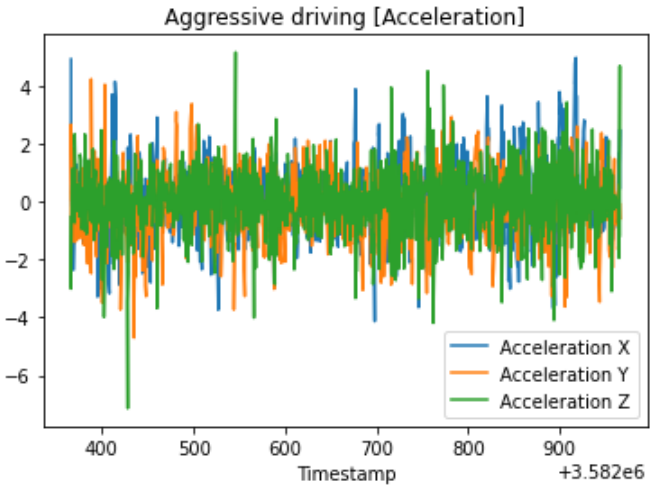

Figure 8. Accelerometer values for aggressive driving

Figure 8 presents the accelerometer values; in this case, we can definitely see a difference between this figure and Figures 4 and 6. We can observe mode accelerations and brakes with one going to -6, and the lateral forces are more present, which indicates sudden accelerations and turns.

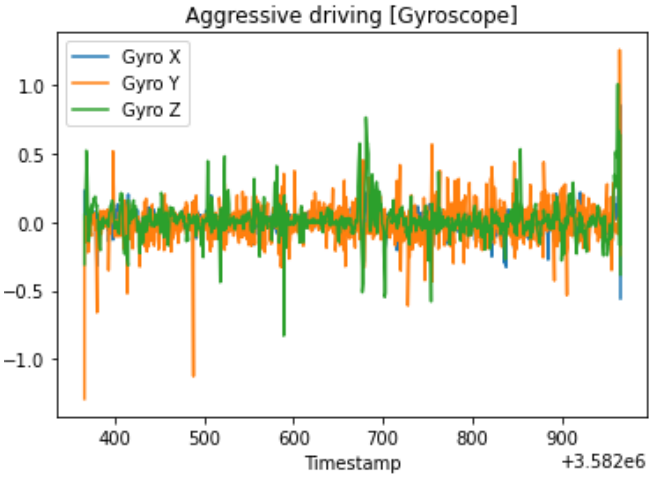

Figure 9. Gyroscope values for aggressive driving

Figure 9 represents the gyroscope values collected for aggressive driving, and we can see here that sudden turns are more frequent with many forces applied on Y axes.

We also won't consider the noise at the beginning and the end of data collecting because we had to move the phone to change the driving style or start/stop the application. Still, overall, the forces applied are more consistent, and we can see how this makes the difference between this figure and Figures 7 and 5, which are more similar than this one. Another thing that needs to be mentioned is that there is a more clear difference between accelerometer and gyroscope values collected when driving aggressive and the other two driving behaviours than between normal and slow driving.

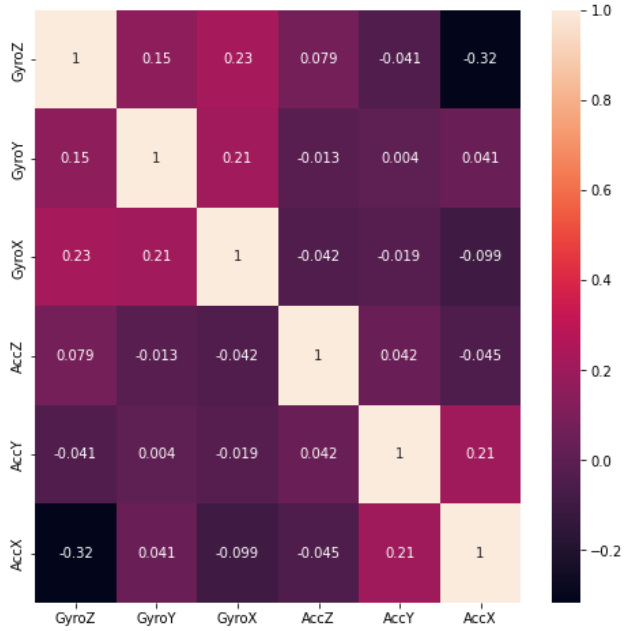

Figure 10. Feature correlation for the training dataset

Figure 10 is a heatmap generated to evaluate the correlation between features. We omitted, in this case, the timestamp and the class, and we chose to keep the values collected from the accelerometer and gyroscope.

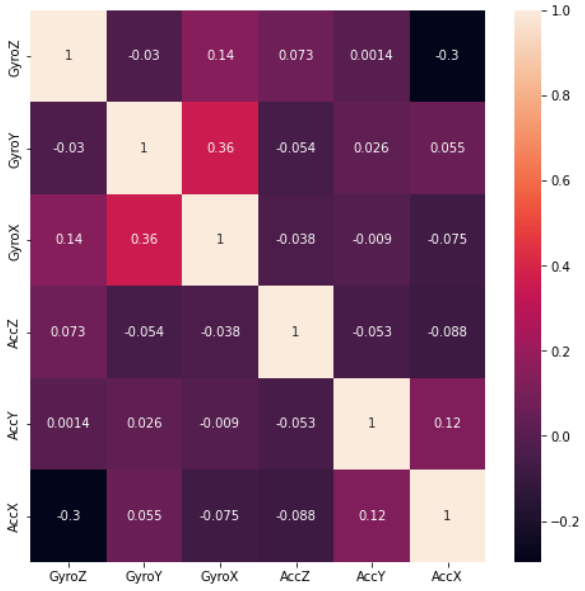

Figure 11. Feature correlation for the test dataset

We chose to add the heatmap for the test dataset, which is presented in Figure 11 because it can provide an overview of the similarities between the collected datasets. Comparing Figure 11 with Figure 10, we can observe that the correlation between features is consistent in these two datasets.

EXPERIMENTS

All the experiments presented in this paper can be reproduced using notebooks available on the code section from Kaggle dedicated page[7].

The motivation for adding the experiments in the paper, even if the focus is on describing the dataset, is that we want to evaluate how accurate can be the models trained and tested on our data. One thing that needs to be mentioned is that even if we can build a variety of models, we need to take into account that driving behaviour is an evolutive approach. We have to consider previous steps when trying to predict it.

The approach to predict the driving style was Random Forest from skLearn and a dense neural network from TensorFlow.

Algorithm	No. of Classes	Accuracy
Random Forest	2	0.591
	3	0.456
XGBoost	2	0.599
	3	0.454
Dense NN	2	0.597
	3	0.461

Table 1. Results for one instance classification

Table 1 presents the results obtained when classifying a single instance based on the sensor's data. It may not be the best solution as driving is an evolutive approach and depends on how long the driver makes sudden moves with the vehicle. However, an average of the classification results may offer a good overview of the driving style. For example, we may have a route of 10 minutes with two samples collected per second; each sample is classified as normal/aggressive or slow/normal/aggressive. At the end of the driving session, we may provide two or three percentages of the driving style, like 40% normal and 60% aggressive or 20% slow, 30% normal and 50% aggressive. A driver can't be in just one class in a city driving environment because the driver's behaviour may also depend on the environment and traffic.

On the other hand, other approaches can use a set of instances or a window included in the training part. This approach is mostly better because it makes the model more noise tolerant; for example, hitting a pothole may result in high sensor values, which will result in a most likely aggressive driving behaviour even if the overall driving is slow. Imagine the situation in which many roads have poor road quality; this would clearly bring results that may not be truly relevant for the real driving style.

Algorithm	No. of Classes	Accuracy
LSTM	2	0.712
	3	0.624
CNN LSTM	2	0.724
	3	0.590
ConvLSTM	2	0.795
	3	0.637

Table 2. Results for a window of instances classification

Table 2 presents the results obtained using LSTM [10] and LSTM derivations networks. Regarding the obtained accuracy, it is slightly better than the results obtained in Table 1, and the results are more relevant because of the window of instances used. For each set of results, we only used *keras_tuner* to find the best model. The first row refers to the standard LSTM network with 2 classes (normal and aggressive) and with three which also adds slow.

The CNN LSTM [11] architecture which is the second algorithm used involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to support sequence prediction. The CNN LSTM model will read subsequences of the main sequence in as blocks, extract features from each block, then allow the LSTM to interpret the features extracted from each block

Regarding the ConvLSTM [12] network presented in the table we need to mention that unlike an LSTM that reads the data in directly in order to calculate internal state and state transitions, and unlike the CNN LSTM that is interpreting the output from CNN models, the ConvLSTM is using convolutions directly as part of reading input into the LSTM units themselves.

CONCLUSION

The dataset presented in the paper is a ready-to use dataset which can be easily used in a machine learning algorithm or in a neural network. All the experiments can be reproduced as the code is available on Kaggle's dedicated page. The popularity of the dataset increased constantly since it was published, and this indicate its usefulness and the interest in such a dataset. The experiments presented in previous section of the paper reveal that the dataset can provide good results even if all of them can be further improved for better accuracy. One particularity of the dataset that makes the experiments more robust and reliable is that it is composed of two parts: train and test, allowing users to validate their models with unseen data.

In future work in this direction, we need to mention that the dataset can be further improved with more quantitative data

[7] https://www.kaggle.com/datasets/outofskills/driving-behavior/code

and also with more various data from different cars, from other drivers and different routes.

REFERENCES

1. Bifulco, G. N., Pariota, L., Brackstione, M., & Mcdonald, M. (2013). Driving behaviour models enabling the simulation of Advanced Driving Assistance Systems: revisiting the Action Point paradigm. Transportation Research Part C: Emerging Technologies, 36, 352-366.
2. Dimitrakopoulos, G., & Demestichas, P. (2010). Intelligent transportation systems. IEEE Vehicular Technology Magazine, 5(1), 77-84.
3. Wawage, P., & Deshpande, Y. (2022). Smartphone Sensor Dataset for Driver Behavior Analysis. Data in Brief, 41, 107992.
4. Romera, E., Bergasa, L. M., & Arroyo, R. (2016, November). Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (pp. 387-392). IEEE..
5. Guo, F. (2019). Statistical methods for naturalistic driving studies. Annual review of statistics and its application, 6, 309-328.
6. Ramanishka, V., Chen, Y. T., Misu, T., & Saenko, K. (2018). Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7699-7707).
7. Wang, H., Wang, X., Han, J., Xiang, H., Li, H., Zhang, Y., & Li, S. (2022). A recognition method of aggressive driving behavior based on ensemble learning. Sensors, 22(2), 644.
8. Ellison, A. B., Greaves, S. P., & Bliemer, M. C. (2015). Driver behaviour profiles for road safety analysis. Accident Analysis & Prevention, 76, 118-132.
9. Carballo, A., Lambert, J., Monrroy, A., Wong, D., Narksri, P., Kitsukawa, Y., ... & Takeda, K. (2020, October). LIBRE: The multiple 3d lidar dataset. In 2020 IEEE Intelligent Vehicles Symposium (IV) (pp. 1094-1101). IEEE.
10. Han, Q., Hu, X., He, S., Zeng, L., Ye, L., & Yuan, X. (2018, November). Evaluate good bus driving behavior with lstm. In International Conference on Internet of Vehicles (pp. 122-132). Springer, Cham.
11. Patel, E., & Kushwaha, D. S. (2022). A hybrid CNN-LSTM model for predicting server load in cloud computing. The Journal of Supercomputing, 78(8), 1-30.
12. Chen, X., Xie, X., & Teng, D. (2020, June). Short-term traffic flow prediction based on convlstm model. In 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC) (pp. 846-850). IEEE.