

Applying Machine Learning for Improving Performance Classification on Driving Behavior

Ahmad Iwan Fadli¹, Selo Sulisty², Sigit Basuki Wibowo³

Abstract—Driving accidents are serious events that could cause fatality. According to WHO's reports, reckless driving behaviors such as speeding, driving under influence, and operating phones while driving are among the main factors that could reduce the focus of drivers while driving. Driving accidents are also difficult to handle on a large scale in a country. Using machine learning classification method to determine whether a driver is driving safely or not can help reduce the risk of driving accidents. Drivers tend to be more careful when they know that their driving behaviors are being monitored. We created a classifier model that can be applied to detection systems to classify whether a driver is driving safely or not safely using travel sensor data, which includes gyroscope, accelerometer, and GPS. The classification methods used in this study are Random Forest (RF) classification algorithm, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). This study shows that RF has the best performance with 98% accuracy, 98% precision, and sensitivity 97%. Performance testing shows that the proposed pre-processing method can increase the classifier sensitivity value in the research dataset. It is hoped that the classifier model can be applied to the driving detection system so that it can reduce the risk of traffic accidents.

Keywords—Driving Behavior, Classification Algorithms, Machine Learning, Gyroscope Data, Accelerometer Data, GPS Data.

I. INTRODUCTION

Driving behavior has an impact on traffic safety, fuel/energy consumption, and gas emissions. Among them, traffic safety has the most concern in every country. According to the World Health Organization (WHO), it was reported that as of 2018, the number of deaths due to traffic accidents reached 1.35 million people worldwide, particularly affecting those living in developing countries [1]. Driving behavior, such as speeding, drunk driving, and cellphone usage, are the main factors that cause lack of attention. Studies show that when drivers are safer when they are monitored. Thus, in order to reduce road accidents, various technologies have been developed to detect the driver's behavior [2]. Traffic accidents do not only have an impact on material losses but also on loss of life. Therefore, it is necessary to take appropriate measures to implement the detection of driving behavior for reducing the risk of accidents to reduce the number of traffic accidents.

Intelligent Transport System (ITS) was first investigated in early 1970s. The primary purpose of ITS is the application of advanced technology on transportation facilities to be more

informative, current, smooth and safe [3]. Fundamental information of ITS, such as driving behavior [4] vehicle mode detection, can be performed by utilizing travel data such as electronic sensors and data transmission. Some sensor data that can be used in this study include GPS data, a combination of accelerometer sensor and gyroscope sensor using statistical methods, and the application of machine learning.

Machine learning is the process of extracting information on a dataset to gain knowledge using specific techniques [5]. The dataset used is a traveling dataset specifically driving data using a car by processing the accelerometer, gyroscope, and speed sensor. Classification technique was performed on data that have been labeled [6]. Previous related studies have researched driving behavior [2], [7], [8]. From several studies that have been conducted, most researchers used gyroscope sensors, while the methods used were SVM, random forest, decision tree, fuzzy, etc.

The research dataset used is travel data with varying mileage randomly drawn from online taxi transaction data operating in Indonesia. The contents of the dataset are the accelerometer sensor data recording, the gyroscope sensor, and the global positioning system (GPS) sensor embedded on the driver's smartphone. The recording process in one trip starts from the pick-up point to the delivery point according to distance, route, and travel time for each trip. The process of labeling safe and dangerous trips is done by looking at reviews and travel ratings from passengers.

The initial stage is data analysis using statistical science to see the characteristics of the dataset, from the process, obtained information that the research dataset has two class labels, "class 0" is a safe trip, and "class 1" is a dangerous trip. The dataset label has a greater amount of safe travel data than hazardous trips in a ratio of 3:1. Imbalance dataset becomes a problem in the classification process because classifiers will be more likely to detect data into dominant classes [9]. Imbalance dataset conditions used in this study resulted in the classifier tending to recognize dangerous trips as safe trips, so the sensitivity value of dangerous trips was low. The low value of sensitivity results in very poor classifier performance, many faults occur in the detection of travel data.

Several studies have approached to overcome the imbalance dataset problem, Reference [10], in the research the researcher expressed that there were several approaches undertaken to overcome the imbalance dataset, one of which was resampling the dataset. Resampling dataset is the process of data duplication or data synthesis based on a minority or majority data target, several methods can be used for resampling, one of which is the random sampling method. Random sampling is very suitable for use with datasets with overlapping data characteristics and has a large amount of data [11]. The random

^{1,2,3} Department of Electrical and Information Engineering, Faculty of Engineering, Universitas Gadjah Mada. Jln. Grafika No. 2, Kampus UGM, Yogyakarta, 55281, Indonesia (Tel.+62-274-552305, email: ahmad.iwan.f@ugm.ac.id; selo@ugm.ac.id; sigitbw@ugm.ac.id)

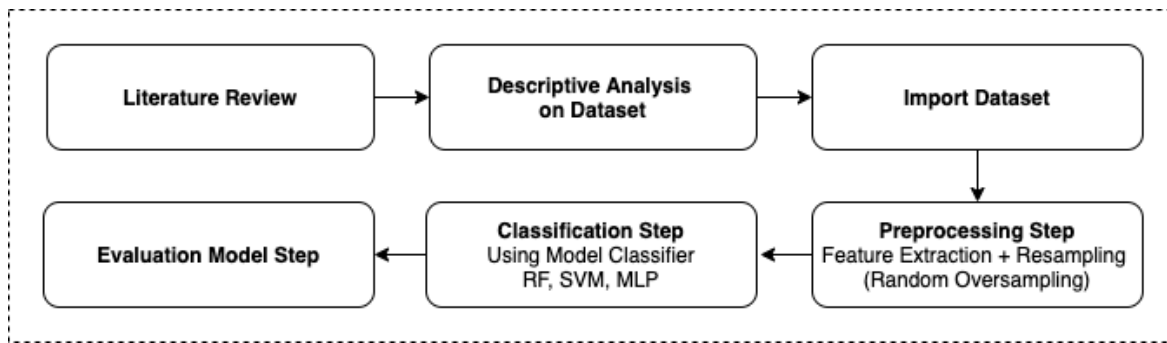


Fig. 1 Research method.

TABLE I
LITERATURE REVIEW

	2017	2017	2018	2015
Topic	Driver behavior profiling [7]	Abnormal driving behaviors [8]	Human activity [9]	Transportation mode [12]
Methods	Classification	Classification	Classification + oversampling	Classification + feature extraction
Algorithms	BN, ANN, RF, SVM	SVM, NN	DT, SVM, MLP	RF, DT, Bootstrap, SVM, KNN
Sensor	Accelerometer, magnetometer, gyroscope	Accelerometer, gyroscope	Accelerometer, audio	Accelerometer, gyroscope, light sensors

sampling method has two techniques that can be used in its application, namely Random Under Sampling (RUS) and Random Over Sampling (ROS). RUS has the advantage of a fast resampling process, but there may be some important information that is lost during the process of erasing data that is close together. ROS has the advantage that the dataset information remains intact, but the length of the resampling process and the size of the dataset will be greater so that the classification process will require longer processing time.

Based on those problems, this research study will carry out the pre-processing data analysis stage, by applying the random oversampling method to overcome the imbalance dataset problem then the feature extraction method is applied to extract the 3-axis feature into a vector. Reference [10] indicates that feature extraction can speed up the process of retrieving data so that the expected results of the method proposed in this study can improve the performance classifier.

In addition to the proposed pre-process data method, we also conducted a test scheme for the classifier that has the best performance in processing gyroscope sensor data and accelerometer in previous studies. Some classifiers include Random Forest (RF) in driver behavior profiling research [7], Multi-Layer Perceptron (MLP) in the imbalanced datasets for human activity [9], and Support Vector Machine (SVM) in transportation mode recognition research [12]. Based on the

characteristics of the dataset used in this study, in general, the dataset has a large amount of data, varied features, and binary label classification so RF and MLP are used because can overcome noise and missing values and are reliable in classifying large datasets, while SVM is very well applied to datasets with binary classification characteristics, SVM can determine hyperplane by selecting fields with optimal margins.

Based on the analysis of the dataset and literature review, this research will propose a pre-processing data method that will be tested on classifiers RF, MLP, and SVM. The purpose of this study is to find the classifier with the best performance in classifying travel data, as an initial step in implementing a driving safety detection system in an online transportation application to reduce the risk of traffic accidents.

II. METHODOLOGY

This paper is divided into several stages consisting of a literature study, descriptive analysis of travel data, experimental stages, and the testing stage by measuring the results of the confusion matrix. The experiments aim to find the best classification method by comparing various classification algorithms.

Fig. 1 shows the research method to be carried out in this study, an experimental scenario will be carried out, namely, an experiment using a method of data pre-processing which will be tested on classifiers RF, SVM, and MLP. The steps undertaken in this study are divided into the following stages.

A. Literature Review

At this stage, a literature review is carried out to study the classification methods that have been used in previous studies. There are many ways to detect driving behavior, depending on the dataset provided in the case. Several similar studies on the detection of driving activity based on relevant research sensor data are discussed as follows.

Table I is some of the studies referenced in this study. In that study, researchers have tested several classifiers and produced the best performing classifiers in classifying the smartphone sensor data. Based on the review of this research library will use classifier RF, SVM, and MLP.

B. Descriptive Analysis on Dataset

The material used in this study is a dataset in the form of sensor data that has been specially installed to record travel data

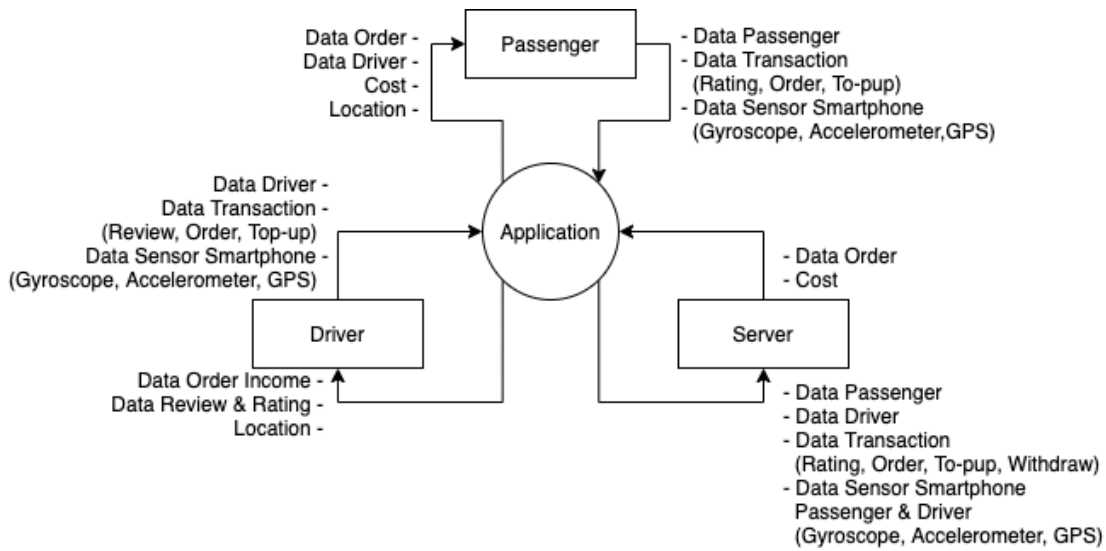


Fig. 2 Diagram context collections dataset.

based on trips with varied mileage and how to drive is taken randomly with the following details.

1. Dataset provider is one of the technology companies operating in Indonesia to serve transportation through online transportation booking services. Fig. 2 shows the diagram context collections dataset in this study. Driver and passenger sensor data will automatically be sent and stored on the application server. The smartphone used based on the standard application usage is Android 4.1 or higher. The application used is an application that has been provided by service providers for drivers and passengers, researchers are not given complete information related to the application due to company privacy.
2. Smartphone sensor data comes from devices embedded in smartphones such as gyroscope MEMS, accelerometer MEMS, and GPS device.
3. The vehicle model used is a car.
4. The data contains a total of about 16 million rows of data.
5. Data labels are available with a total of 20,018 of data. This data contains Booking IDs that have been labeled.
6. The research dataset obtained already has a label. not a lot of information that researchers get related to the process of labeling the dataset due to company data privacy issues, but the labeling process is done by filtering the trip based on travel reviews and validated by the company's transportation experts by looking at gyroscope, accelerometer, and speed data statistics. Table II shows the primary information available on each dataset provided. The table shows that there are 11 main features and labels with two classes.

C. Pre-processing Step

The raw data processed in this study contains total data of about 16 million rows out of 20,018 trips. The amount of data per flight varies based on the distance traveled, so the process of classifying driving behavior cannot be done on whether the driver is driving safe or dangerous.

TABLE II
DESCRIPTION OF DATASET FEATURE

No	Feature	Descriptions
1	BookingID	Travel ID
2	Bearing	Compass direction
3	Accuracy	Accuracy measurement based on GPS
4	Acceleration_x	Measurement of accelerometer on the x-axis
5	Acceleration_y	Measurement of accelerometer on the y-axis
6	Acceleration_x	Measurement of accelerometer on the z-axis
7	Gyro_x	Measurements of gyroscope on the x-axis
8	Gyro_y	Measurements of gyroscope on the y-axis
9	Gyro_x	Measurements of gyroscope on the z-axis
10	Second	Record time of data recording
11	Speed	Speed measurement
12	Label	- Class '0' Safe trip - Class '1' Dangerous trip

Fig. 3 is the pre-process dataset stage. At the stages, there are several activities carried out to process the data so that it is ready to be processed by the classifier. Some steps in the pre-processing data are as follows.

1) *Import Dataset*: This process is the initial stage carried out at the pre-processing step. This research dataset has two types of data files, feature data files, and label data files. Because the amount of data is large enough so that the data feature is divided into several parts, therefore it is necessary to merge the data parts into one feature data section.

2) *Feature Extraction*: The next stage is the feature extraction stage. Information obtained at this stage that in 16,135,561 data rows provided only 83,146 data rows have labels. From some existing features, it is extracted into a new

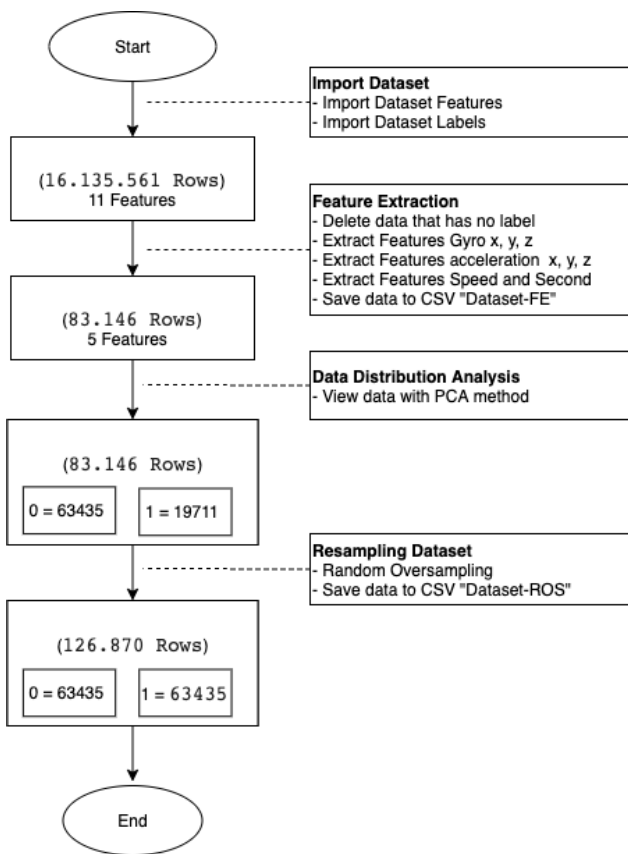


Fig. 3 Pre-processing step.

feature. Before going through the classification stage, the feature extraction method applied to the accelerometer sensor data x, y, z, gyroscope x, y, z, and data GPS.

Fig. 4 is the feature extraction process, after process the dataset will be saved into a CSV file for the method testing process. These features will be used by the classifier engine to distinguish two different types of class objects. The result of feature extraction will be through mapping into two kinds of safe and dangerous directories. As mentioned earlier, mapping the type of movement into these two types of lists will facilitate the labeling process in the feature extraction stage. Any data that has gone through the feature extraction process will be able to find the appropriate travel label easily.

3) *Resampling Dataset*: The following data pre-processing process is resampling the dataset. Resampling method proposed because dataset imbalance, also described in the label data, is that class 0 is a safe trip, and class 1 is a dangerous trip. The amount of safe travel data is greater when compared to dangerous trips in the ratio of 3:1. Several approaches used to overcome the imbalance dataset, one of which is resampling the dataset. The resampling dataset is a process of data duplication or data synthesis based on a minority or majority data target.

Fig. 5 is the resampling process that will be applied, the dataset after feature extraction (written as FE in Fig. 5) will be processed using the random oversampling method. After the resampling process is complete the dataset will be saved into the dataset-ROS

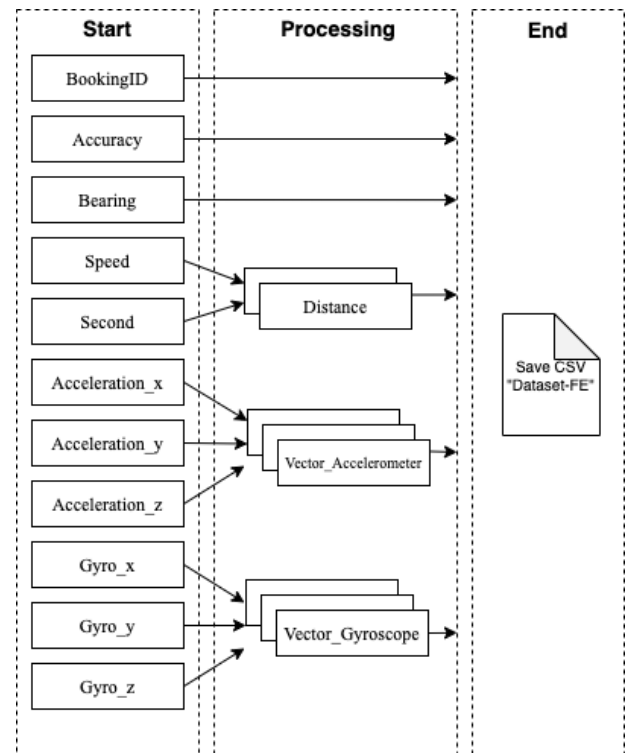


Fig. 4 Scheme of feature extraction.

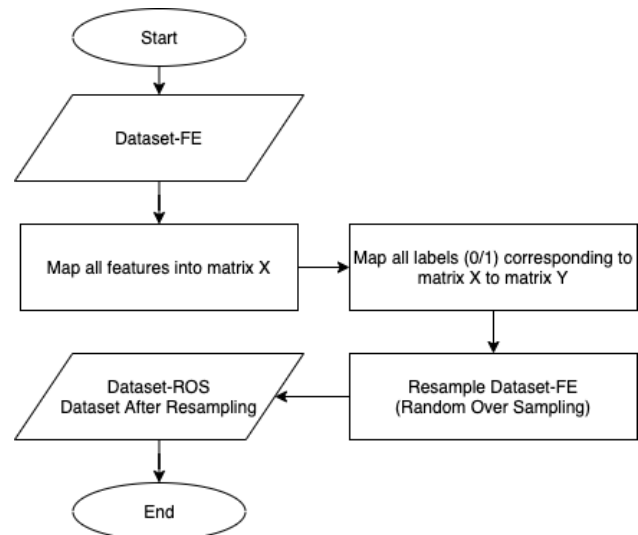


Fig. 5 Scheme of resampling dataset.

D. Classification Step

After going through the pre-process data, the final dataset is obtained five features of three smartphone sensors have a high significance in distinguishing the safe and dangerous travel classes. The five features derived are Accuracy, Bearing, Vector_gyroscope, Vector_Acceleration, and Distance.

The classification methods used in this study are the RF, SVM, and MLP methods. Based on observations of data distribution using the Principal Component Analysis (PCA) method, the characteristics of data distribution are that data points tend to overlap each other. SVM is used as a classification method because it can form clear lines separating

TABLE III
CONFUSION MATRIX FOR DRIVING BEHAVIOR

	Predicted Safe	Predicted Dangerous
Actual safe	True Positive (TP)	False Negative (FN)
Actual dangerous	False Positive (FT)	True Negative (TN)

TABLE IV
STATISTICAL DATA DESCRIPTION

Parameter	Gyroscope	Accelerometer	GPS Speed
Mean	0.1293312	9.893247	9.006543
Median	0.0667378	9.817840	7.5299997
Std Dev	37.64916	1.268255	8.10641
Minimum	0	0.0050742	0
Maximum	80.77798	115.15224	148.0186
Count	16,135,561	16,135,561	16,135,561

overlapping objects. Similar to SVM, RF is a classification method that can separate two overlapping objects, but the RF method is considered better to do in large datasets as well as MLP. These three classification methods are also used in previous research experimental models.

E. Evaluation Model Step

The testing scenario that will be carried out is as follows.

1. Testing the imbalance dataset using classifiers RF, SVM, and MLP.
2. Testing feature extraction FE in the imbalance dataset using classifiers RF, SVM, and MLP.
3. Testing the ROS in the imbalance dataset using classifiers RF, SVM, and MLP.
4. Testing combines feature extraction, and ROS in the imbalanced dataset using classifiers RF, SVM, and MLP.

After four experimental schemes classification methods using RF, SVM, and MLP classification methods are performed. Furthermore, the researcher will complete the validation process, the validation employed in this research is the split-validation method. The split-validation has a lower variance when compared to split training-test data, and it indicates that the data points will be likely to be close to the Expected value. The small number of variances will be relevant to the circumstances in which the data is limited. The result of this validation process is the Confusion matrix. From the results, the confusion matrix can calculate the performance of classification models such as accuracy, precision, and sensitivity.

Accuracy is used to determine how much the proportion of an instance of a test is classified correctly (positive and negative) on all data. Precision is used to determine the correct positive prediction ratio compared to the overall predicted positive results. Sensitivity is used to determine how much the proportion of the test instance is correctly classified on a safe journey. The three parameters used in this study were from the confusion matrix in Table III. The table shows the data classification assessment component described as follows.

1. True positive indicates the number of instances identified correctly on a safe trip.

2. False positive indicates the number of instances identified incorrectly on a safe trip.
3. True negative shows the number of instances that were correctly identified on dangerous trips.
4. False negative indicates the amount of data incorrectly identified on a dangerous journey.

III. RESULTS AND ANALYSIS

This study has two main stages, namely (A) Pre-processing data Steps, and (B) Classification Steps. The data pre-processing stage begins with the application of the feature extraction method, then continues with the resampling method, and lastly to find out what features are most influential using the feature selection method.

A. Analysis of Pre-processing Steps

This pre-processing stage has several stages in which the stages are as follows.

1) *Dataset Characteristics*: This study uses two datasets provided, namely the feature dataset and the label dataset, the feature dataset contains gyroscope sensor data recording data, acceleration sensor, and GPS Sensor, while the label dataset is a BookingID dataset that has been labeled safe and dangerous travel. the characteristics of the two datasets if described statistically it will look like in Table IV.

Table IV is an analysis table for the dataset in the table. It is known that the main features used in this study are the accelerometer sensor, gyroscope, and GPS data. Next, the sensor data available in this research dataset are visualized.

Fig. 6 is a visualization data from the accelerometer sensor data. The accelerometer sensor is the result of the Acc xyz vector for time. Data visualization above illustrates the movement of the accelerometer sensor on safe trips and dangerous trips.

Fig. 7 is the result of the visualization of the gyroscope sensor data. Gyroscope data visualization is the result of the xyz gyro vector for time. Gyroscope data that we show is a trip sampling that already has a label. The data above shows the gyroscope movement on dangerous trips is relatively unstable compared to safe travel.

Fig. 8 is the result of the visualization of GPS data. GPS data mapping used is time and speed data. The GPS data we show is a sampling of trips that already have a label. From the data above shows, the speed on dangerous trips is relatively higher than safe travel.

2) *Analysis of Feature Extraction Stage*: Gyroscope sensor, accelerometer, and GPS have several features in it. From these features a new feature can be extracted. The feature extraction process uses a Python script by adopting a 3-dimensional vector equation that is applied to the gyroscope sensor and the accelerometer sensor. Then, for GPS features, new features can be extracted from speed and time. The equation used is as follows.

$$Vector_{Gyroscope} = \sqrt{Gyro_x^2 + Gyro_y^2 + Gyro_z^2}. \quad (1)$$

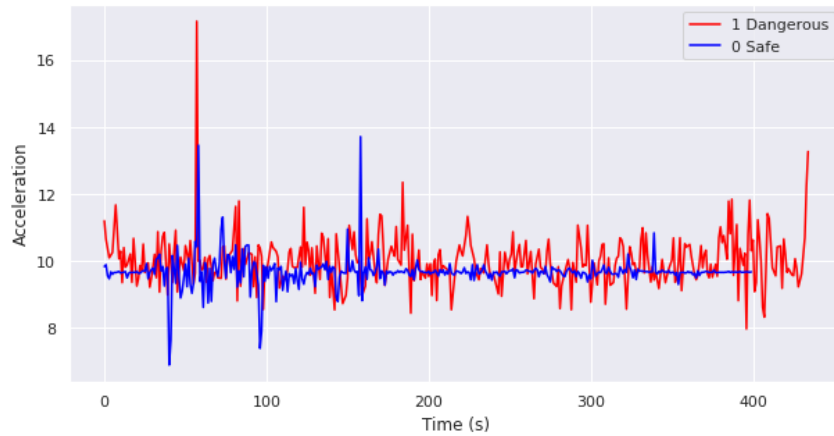


Fig. 6 Visualization of accelerometer sensor data.

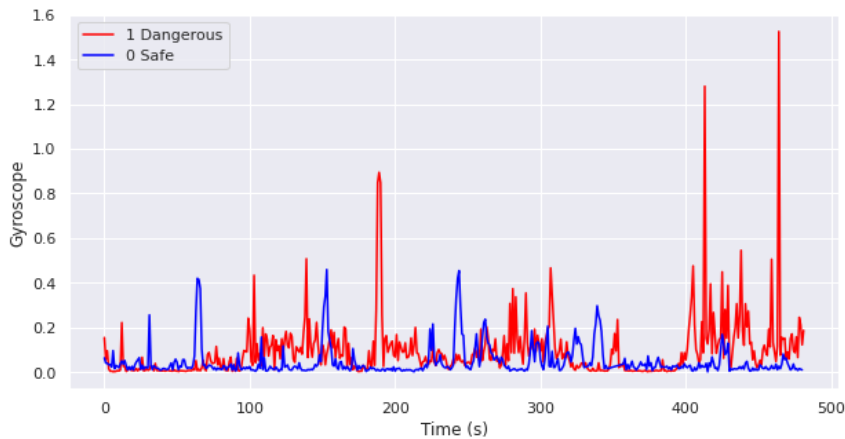


Fig. 7 Visualization of gyroscope sensor data.

Equation (1) is a formula for obtaining new features by utilizing gyroscope sensor data such as gyro x, gyro y, and gyro z data, from which all three data can be found a vector point. The new feature generated from the Extraction Feature using the script above is the Vector_Gyroscope feature.

$$Vector_{Accelerometer} = \sqrt{Acc_x^2 + Acc_y^2 + Acc_z^2}. \quad (2)$$

Equation (2) is a formula to get new features by using accelerometer sensor data such as data acc x, acc y, and acc z, from these three data, a vector point can be determined. New features generated from the extraction features using the above script are the Vector_Accelerometer feature.

$$Distance = Speed \times Second. \quad (3)$$

Equation (3) is a formula to get new features by utilizing GPS data such as speed and time data, from the two data, a distance can be created in seconds. New features generated from the extraction feature using the script above are the distance feature.

From the feature extraction process, we get three new datasets from the original dataset, after going through the feature extraction process, the final features obtained are as follows as in Table V.

TABLE V
FEATURE EXTRACTION RESULTS

No	Feature	Amount of Data	Obtained from
1	Accuracy	83,146 rows	Original feature
2	Bearing	83,146 rows	Original feature
3	Vector_Accelerometer	83,146 rows	Feature extraction
4	Vector_Gyroscope	83,146 rows	Feature extraction
5	Distance	83,146 rows	Feature extraction

3) *Analysis of the Oversampling Phase:* The feature extraction process produces a new dataset with an insane amount of data between safe and dangerous movements. Data with safe motion classes amounted to 63,435 rows, while for dangerous movement classes numbered 19,711 data. The ROS resampling method is used to balance the data so that the classification engine does not lean towards classes that have more numbers. Following is the resampling process using the random oversampling method. The following is an illustration of the condition of the dataset before and after going through the resampling stage.

Fig. 9 is a description before and after the resampling process. After checking the count of data it is known that the

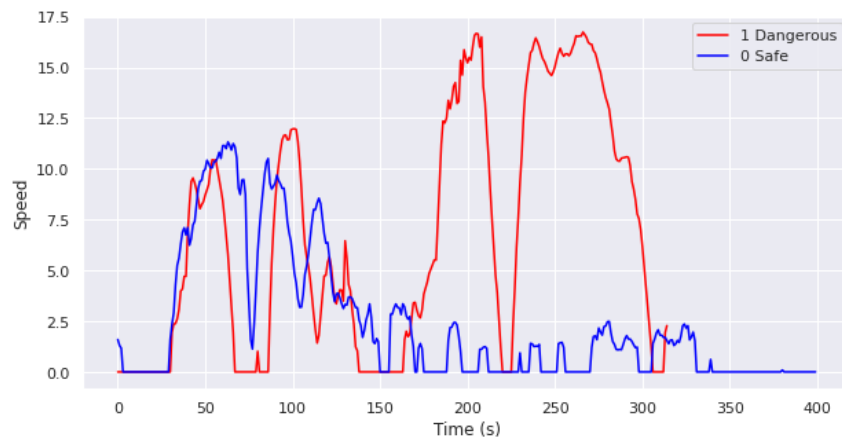


Fig. 8 Visualization of GPS data.

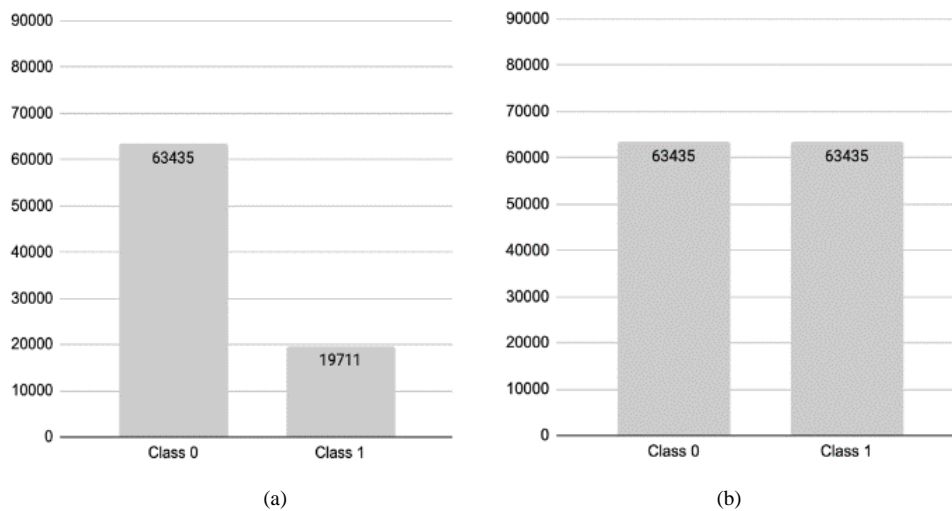


Fig. 9 Description of the resampling dataset, (a) before resampling, (b) after resampling.

dataset is not balanced between safe and dangerous travel labels. Data with safe motion classes amounted to 63,435 rows, while for dangerous movement classes numbered 19,711 data. The ROS resampling method is used to balance the data so that the classification engine does not lean towards types that have more numbers.

Table VI is a list generated at the oversampling stage. Changes that occur in the oversampling process is the amount of data, before oversampling the amount of data is 83,146 rows after the oversampling process becomes 126,870 rows

B. Analysis of Classification Stage

In the classification process, the results of the classification process are as follows.

1) *Random Forest (RF)*: The following are the results of testing the RF classifier against several proposed test scenarios. The parameter settings used of random forest classifier are shown in Table VII.

The test will use several different models from the imbalance dataset to the application of feature extraction and random oversampling, the results of testing in the RF classifier are

TABLE VI
OVERSAMPLING RESULTS

No	Feature	Amount of Data	Obtained from
1	Accuracy	126,870 Rows	ROS
2	Bearing	126,870 Rows	ROS
3	Vector_Accelerometer	126,870 Rows	ROS
4	Vector_Gyroscope	126,870 Rows	ROS
5	Distance	126,870 Rows	ROS

shown in Fig. 10. The first test graph shows that RF has fairly good accuracy and precision value, in handling imbalance data with a performance value above 90%, but the resulting sensitivity is quite low when seen from the accuracy results obtained.

The second test applies the feature extraction method into the imbalanced dataset. The results of the RF performance test show a slight increase in the value of accuracy and precision, but a very drastic increase in the sensitivity value of 28%. The performance improvement is caused by the feature extraction method by simplifying 11 features into five features, so that the tree selection process is shorter.

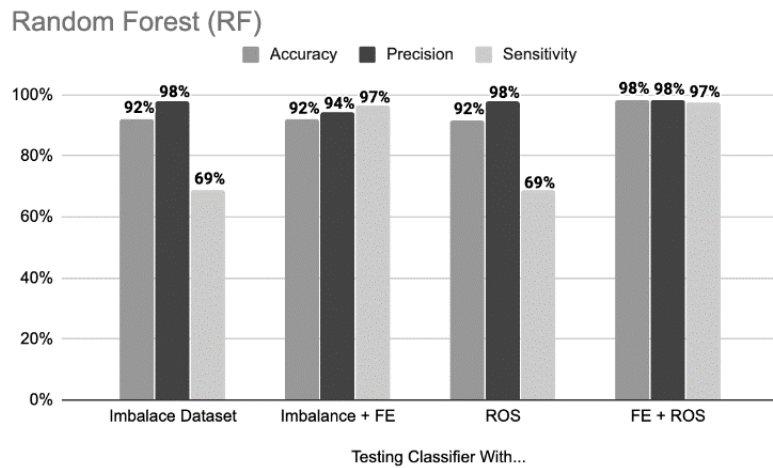


Fig. 10 Result of testing in random forest classifier.

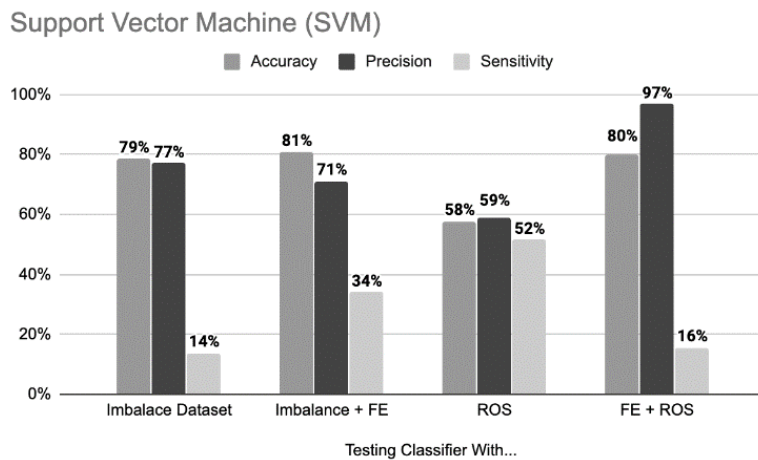


Fig. 11 Result of testing SVM.

TABLE VII
PARAMETER OF RANDOM FOREST

No	Parameter	Value
1	model_selection	train_test_split
2	train_size	0.7
3	test_size	0.3
4	bootstrap	True
5	ccp_alpha	0
6	criterion	gini
7	min_impurity_decrease	0
9	min_samples_leaf	1
10	min_samples_split	2
11	min_weight_fraction_leaf	0
12	n_estimators	100
13	oob_score	False
14	verbose	0
15	warm_start	False

TABLE VIII
PARAMETER OF SVM

No	Parameter	Value
1	model_selection	train_test_split
2	train_size	0.7
3	test_size	0.3
4	C	1
5	break_ties	False
6	cache_size	200
7	decision_function_shape	ovr
8	degree	3
9	gamma	scale
10	kernel	rbf
11	max_iter	-1
12	probability	False
13	random_state	None
14	shrinking	True
15	tol	0.001
16	verbose	False

The third test applies the ROS method. The test results show an increase in RF performance on the precision value, but there is a decrease back in the value of sensitivity, a decrease in performance is due to RF is good in overcoming Imbalance data so that the oversampling process does not affect the RF classifier.

The fourth test carried out a combination of the feature extraction method and ROS. The test results showed sufficient performance improvements to the value of accuracy and precision from the first to the fourth test, but there was an

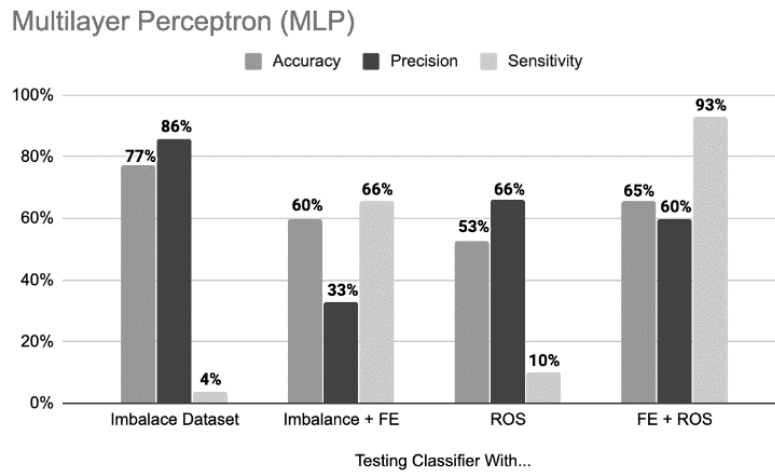


Fig. 12 Result of testing MLP.

increase in the sensitivity value again. The increased sensitivity is due to the Feature extraction method which simplifies the tree selection process. The fourth test has the best RF performance value of all tests performed with an accuracy value of 98%, precision 98%, and sensitivity 97%.

2) *Support Vector Machine (SVM)*: The following are the results of testing the SVM classifier against several proposed test scenarios. The parameter settings used of SVM classifier are shown in Table VIII. The test uses several different models from the imbalance dataset to the application of feature extraction and random oversampling, the results of testing in the SVM classifier are as follows.

Fig. 11 is the result of testing SVM of classification stage. The first test graph shows that SVM has fairly good accuracy and precision value, in handling imbalance data, but the resulting sensitivity value is very low.

The second test applies the feature extraction method into the imbalanced dataset, the results of the SVM performance test showed a significant increase, where the sensitivity value was increased by 20%, but there was a decrease in the precision value by 6%. The increase and decrease in performance are caused by the feature extraction method by simplifying 3-axis XYZ features into 1 vector feature so that the matrix created by the classifier is simpler processed by SVM.

The third test applies the ROS method. The test results show a decrease in SVM performance on the value of accuracy and precision, but there is an increase in the sensitivity value. The resulting widening and the classification process more severe. The fourth test carried out a combination of the feature extraction method and ROS test results showed a very high-performance increase in the value of accuracy and precision, but there was also a drastic decrease in the sensitivity value. The decrease in sensitivity value is caused by the process of adding a synthesis data ROS method causing more duplication in the method of feature extraction with the most dominant class affecting even changing the information available in the minority class. However, data duplication has a good effect on SVM accuracy with the highest test values of all tests with an accuracy value of 80% and 96% precision.

TABLE IX
PARAMETER OF MLP

No	Parameter	Value
1	model_selection	train_test_split
2	train_size	0.7
3	test_size	0.3
4	activation	relu
5	alpha	0.0001
6	batch_size	auto
7	beta_1	0.9
8	beta_2	0.999
9	early_stopping	False
10	epsilon	1e-08
11	hidden_layer_sizes	100
12	learning_rate	constant
13	learning_rate_init	0.001
14	max_fun	15,000
15	max_iter	200
16	momentum	0.9
17	random_state	None
18	solver	adam
19	tol	0.0001
20	validation_fraction	0.1

3) *Multilayer Perceptron (MLP)*: The following are the results of testing the MLP classifier against several proposed test scenarios. The parameter settings used of MLP classifier are shown in Table IX. The test will use several different models from the imbalance dataset to the application of feature extraction and random oversampling, the results of testing in the MLP classifier are shown in Fig. 12.

The following are the results of testing the MLP classifier against several proposed test scenarios. The test uses several different models from the imbalance dataset to the application of feature extraction and random oversampling, the results of testing in the MLP classifier are as follows.

The first test graph shows that MLP has fairly good accuracy and precision value, in handling imbalance data with an accuracy value of 77% and 85% precision, but the resulting sensitivity is very low when viewed from the accuracy results

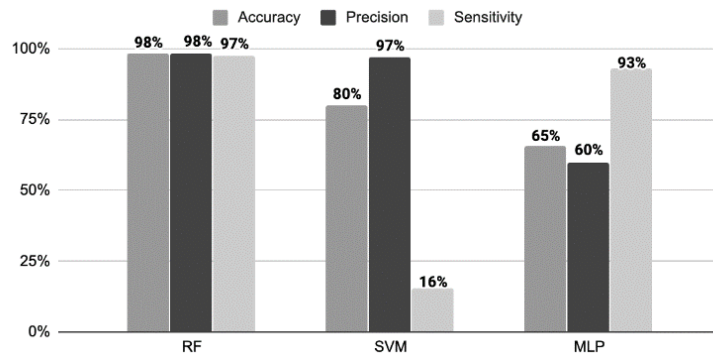


Fig. 13 Results classifier comparison.

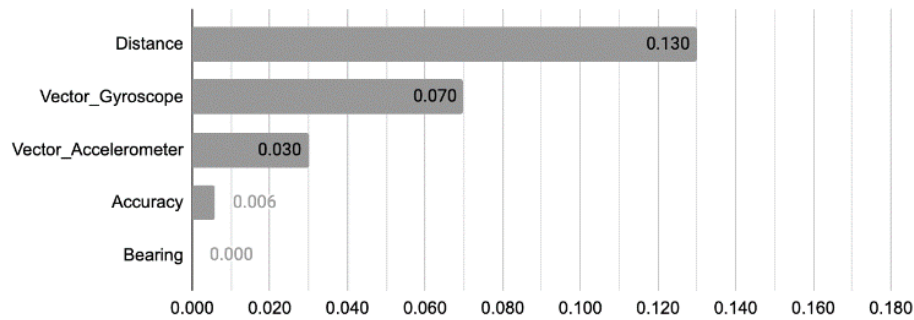


Fig. 14 Results of most influential feature.

obtained. The second test applies the feature extraction method to the imbalance dataset, the results of the MLP performance test show a very drastic decrease in performance at an accuracy value of 17% and a precision of 53%, but a very drastic increase in the sensitivity value of 62%.

The third test applies the ROS method. The test results show a decrease in MLP performance on the value of accuracy and sensitivity, the decrease in performance is due to MLP being good at overcoming Imbalance data so that the oversampling process does not really affect the MLP classifier.

The fourth test carried out a combination of the feature extraction method and ROS test results showed a decrease in performance of the accuracy and precision of imbalance data, but there was a very drastic increase in the sensitivity value. The fourth test had a more balanced MLP performance value of all the tests carried out 65% accuracy, 60% precision, and 92% sensitivity.

C. Analysis of Classifier Models Comparison Stage

After going through the stages of classifier testing using the proposed methods such as feature extraction and ROS, we compared the performance of RF, SVM and MLP classifiers using the results of the matrix confusion of each classifier, following the results obtained.

Table X is the confusion matrix of RF. From the results, RF is a good-fitting model because it has few false in predictions so that the resulting high sensitivity and accuracy values, the model built can explain data without being affected by data noise.

Table XI is the confusion matrix of SVM. From the results, SVM is an under-fitting model because it has many false in

TABLE X
CONFUSION MATRIX OF RANDOM FOREST

RF	Predicted Safe	Predicted Dangerous
Actual safe	18,792	265
Actual dangerous	354	18,650

TABLE XI
CONFUSION MATRIX OF SVM

SVM	Predicted Safe	Predicted Dangerous
Actual Safe	12,187	9,806
Actual dangerous	9,198	6,870

TABLE XII
CONFUSION MATRIX OF MLP

MLP	Predicted Safe	Predicted Dangerous
Actual safe	7,244	11,813
Actual dangerous	1,338	17,666

predictions so that the resulting sensitivity and accuracy values are low, this happens because the model cannot see the logic behind the data, so it cannot make predictions correctly, both for training datasets and other similar datasets.

Table XII is the confusion matrix of MLP, from the results MLP can predict dangerous trips very well, but also many false predicted on trips that should be safe, this method tends to predict all data into dangerous trips. This condition causes high sensitivity, but the low accuracy value causes the model to over-fitting because the model is made too focused on the training of a particular dataset, so it cannot make predictions correctly if given another similar dataset. MLP model tends to capture noise data that should be ignored.

From the test scheme conducted on the research classifier using the proposed method, Fig. 13 shows a comparison between RF, SVM, and MLP. RF has the best performance, which is 98% accuracy, 98% precision, and 97% sensitivity. The best model performance results from a combination of feature extraction and random oversampling.

D. Analysis of the Most Influential Feature

After going through the classifier stage, what features are most influential on driving detection are also found. The ranking of the most prominent features is shown in Fig. 14. The graph shows what features are most influential in the classification process. The test results show that distance is the most prominent data in determining how to drive safe or dangerous categories.

IV. CONCLUSIONS

This study classifies travel datasets by processing smartphone driver sensor data, based on testing using RF, SVM, and MLP classifiers showing good accuracy, but the resulting sensitivity is very low. It is caused by the dataset used imbalance, therefore classifier tend to predict in the dominant class. To overcome this problem, pre-processing needs to be done on the dataset so that the predicted value of true negative increases so that the sensitivity value will also increase. This study proposes the feature extraction method and ROS to overcome these problems. The results of feature extraction testing showed an increase in the RF, SVM, and MLP sensitivity values, then the ROS method testing did not influence the RF and MLP classifiers because the characteristic of these two classifiers is reliable in handling imbalance data, but the ROS method is very influential on the SVM classifier.

From the test scheme conducted on the research classifier using the proposed method between RF, SVM, and MLP. RF has the best performance, which is 98% accuracy, 98% precision, and 97% sensitivity. The best model performance results from a combination of feature extraction and random oversampling.

The results of this study are the first step in the application of artificial intelligence technology in the field of ITS to transportation facilities to be more informative, current, smooth and safe that can overcome traffic safety problems in large scale. In its implementation, the research model can be made as a detection and prediction module that can be applied to the driving safety detection system. With the existence of a safety detection system motorists can provide notification directly to

the driver who indicated driving dangerously to be careful in driving. Then the benefits for law enforcement in the field of transportation can facilitate the process of investigation of a traffic accident based on the results of the analysis produced by the system by utilizing travel data records.

With the existence of a vehicle safety detection system, it is expected to reduce the number of traffic accidents, so that the number of deaths due to accidents can decrease and the burden of a country's losses in bearing accident costs can be allocated to improve a better transportation system in the future.

REFERENCES

- [1] (2018) World Health Organization - WHO website, [Online], <https://www.who.int/publications/i/item/global-status-report-on-road-safety-2018>, access date: 16-Sep-2019.
- [2] C. Ma, X. Dai, J. Zhu, N. Liu, H. Sun, and M. Liu, "DrivingSense: Dangerous Driving Behavior Identification Based on Smartphone Autocalibration," *Mobile Information Systems*, Vol. 2017, pp. 1–15, Mar. 2017.
- [3] L. Zhu, F.R. Yu, Y. Wang, B. Ning, and T. Tang, "Big Data Analytics in Intelligent Transportation Systems: A Survey," *IEEE Trans. Intell. Transp. Syst.*, Vol. 20, No. 1, pp. 383–398, Jan. 2019.
- [4] D.N. Lu, D.N. Nguyen, T.H. Nguyen, and H.N. Nguyen, "Vehicle Mode and Driving Activity Detection Based on Analyzing Sensor Data of Smartphones," *Sensors (Switzerland)*, Vol. 18, No. 4, pp. 1–25, Mar. 2018.
- [5] X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data Mining with Big Data," *IEEE Trans. Knowl. Data Eng.*, Vol. 26, No. 1, pp. 97–107, Jan. 2014.
- [6] P. Patil, N. Yaligar, and S. Meena, "Comparison of Performance of Classifiers - SVM, RF and ANN in Potato Blight Disease Detection Using Leaf Images," *IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC*, 2017, pp. 1–5.
- [7] J.F. Júnior, E. Carvalho1, B.V. Ferreira1, C.D. Souza, Y. Suhara, A. Pentland, and G. Pessina, "Driver Behavior Profiling: An Investigation with Different Smartphone Sensors and Machine Learning," *PLoS ONE*, Vol. 12, No. 4, pp. 1-16, Apr. 2017.
- [8] J. Yu, Z. Chen, Y. Zhu, Y.J. Chen, L. Kong, and M. Li, "Fine-Grained Abnormal Driving Behaviors Detection and Identification with Smartphones," *IEEE Trans. Mob. Comput.*, Vol. 16, No. 8, pp. 2198–2212, 2017.
- [9] K.T. Nguyen, F. Portet, and C. Garbay, "Dealing with Imbalanced data sets for Human Activity Recognition using Mobile Phone Sensors," *Intelligent Environments.*, Vol. 3, pp 129-138, Jun. 2018.
- [10] H. He and E.A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, Vol. 21, No. 9, pp. 1263–1284, Sep. 2009.
- [11] A. Ghazikhani, H.S. Yazdi, and R. Monsefi, "Class Imbalance Handling Using Wrapper-based Random Oversampling," *ICEE 2012 - 20th Iran. Conf. Electr. Eng.*, 2012, pp. 611–616.
- [12] A. Jahangiri and H.A. Rakha, "Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data," *IEEE Trans. Intell. Transp. Syst.*, Vol. 16, No. 5, pp. 2406–2417, Oct. 2015.