Smartphone Motion Sensor Data Processing for Driving Characteristics Classification

Lisa Dinda Yunita[1], Ema Utami[2], Ainul Yaqin[3]

[1,2,3] Magister of Informatics Engineering Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Sleman 55281 INDONESIA (email: [1]lisadinday@students.amikom.ac.id, [2]ema.u@amikom.ac.id, [3]ainulyaqin@amikom.ac.id)

[Received: 7 March 2023, Revised: 23 June 2023]
Corresponding Author: Lisa Dinda Yunita

ABSTRACT — Driving behavior significantly influences road safety. Unsafe driving behaviors, such as driving under the influence, speeding, and using mobile phones, can lead to serious accidents and fatalities. This research aims to observe driving characteristics by utilizing smartphone motion sensor data. The data collection method involved recording the driver's smartphone motion sensor during trips. The data were then exported from the system for further processing. The main objective of this study is to process the data by creating a classification model with the best performance in handling smartphone motion sensor data. The results of this research are expected to be implementable models to address road safety issues in the future. Additionally, by utilizing driver characteristic detection technology, awareness of safe driving practices can be enhanced. The research methodology used data mining with machine learning classification modeling using random forest (RF), support vector machine (SVM), and decision tree (DT) methods. The test results indicate that the RF model performed the best with an accuracy of 91.22%. Furthermore, this study found that speed was the most influential factor in identifying safe or unsafe driving behavior. The developed classification model shows the potential to improve traffic management efficiency and contribute to safer transportation. By leveraging driver characteristic detection technology, it is hoped that awareness of safe driving practices will increase, leading to a safer road environment.

KEYWORDS — Data Mining, CRISP-DM, Driving Behavior, Machine Learning, Classification.

I. INTRODUCTION

Traffic accidents have increased significantly from year to year. People prefer to drive their vehicles due to the limited availability of public transportation, causing an increase in the number of vehicles. This increase may raise the risk of traffic accidents. Based on a report on road safety in each country, the World Health Organization (WHO) points out several factors that cause death [1]. Among these factors, traffic accidents are the eighth leading cause of death. In addition, the data obtained in 2018 showed that more than 1 million traffic accidents were reported [1].

Based on several studies, unsafe driving behavior often include drunk driving, speeding, and using cell phones while driving. These driving behaviors are most often the main factor in losing concentration, resulting in distracted driving. Research on the supervision of driving behavior by supervising drivers has also been conducted. The study's findings indicate that supervised drivers demonstrate safe driving behavior, underscoring the significance of implementing systematic computer monitoring for enhanced road safety. To effectively reduce road accidents, continuous and real-time monitoring is crucial, and this can be achieved through the adoption of systematic computer monitoring systems. Ongoing research is dedicated to developing advanced detection technology to further improve monitoring capabilities and promote safer driving practices. The combination of effective monitoring and advanced detection technology offers great potential to create a safer driving environment and reduce the frequency of accidents. Continuous driver's behavior monitoring can raise awareness and encourage responsible driving practices, thus contributing to the overall safety and efficiency of transportation systems. However, the implementation could have been more optimal since the implementation depends on existing mass applications [2]. In addition to material losses, traffic accidents may cause death. Therefore, an appropriate action is needed to implement driving behavior detection to reduce the risk of accidents in the hope of meeting future global targets that may be set and reducing the number of traffic accidents.

In 1988, the Organization for Economic Cooperation and Development (OECD) in Paris stated that every year countries spent more than 1 billion dollars only for transportation [3]. The intelligent vehicle highway system (IVHS) reported that in the United States in 1991, 41 thousand people died from traffic accidents, and more than 5 million people were injured. The report also stated that traffic accidents annually costed the United States $100 billion. After studying the characteristics of the problems, IVHS became an intelligent transportation system (ITS) [4].

ITS is a theory that is expected to overcome transportation problems by utilizing technology. The utilization of technology aims to improve traffic security, data monitoring, and real-time decision-making systems. In many ways, the ITS can support a distributed mass transportation system [5]. This study began with the transaction and operational monitoring and focused on addressing operational monitoring problems, where a classification model was generated for driving by utilizing trip data. The data needed to realize an ITS technology was motion sensor data on the driver's smartphone. Statistical methods can be used to generate some critical information in ITS, including driving behavior detection information and vehicle mode detection used [6]. Global positioning system (GPS) data, accelerometer, and gyroscope served as research data. Therefore, this study classified safety driving detection to see if the driver was driving normally or aggressively, from which, conclusion could be drawn whether the driver engaged in safe or negligent driving. A highly computational mechanism was required to transform the data into system knowledge and a driving detection parameter during data processing. Hence, machine learning technology can be employed as the method [7].

Machine learning is a computational technique that studies data into information and then forms knowledge. It has several processes, including data import, preprocessing, modeling, and validation. This study implemented machine learning techniques using the classification method. Classification method was selected since based on the characteristic of the data provided, the dataset had a label that could be used as a reference value [8]. The dataset labeling process underwent appropriate processes, from creating labeling rules to labeling execution. In this study, the obtained data had been labeled and no information related to the labelling process was collected. The dataset used was ready to be processed with the .csv extension. Features available in the dataset included booking ID; time; gyroscope x, y, and z; accelerometer sensors x, y, and z; speed data. The overall type of dataset was text data. The machine learning technique is a classification technique by grouping data based on objects that have been manually labeled so that a classification model can detect driving behavior on an upcoming trip [9]. Previous related researchers have studied driving behavior. Researchers of previous study mostly used a gyroscope sensor, while the method used was a classification method by applying several support vector machines (SVM) and random forest (RF) algorithms.

In a previous performance evaluation of machine learning algorithms (MLA), artificial neural networks (ANN), SVM, RF, and Bayesian networks (BN) were employed to detect driving behavior using three smartphone sensors [10]. Meanwhile, the features used were four dataset features: accelerometer, linear accelerometer, magnetometer, and gyroscope. This study collected sampling trips randomly from two different drivers. Based on the results of the tests, gyroscope was the best sensor to detect driving behavior, and RF was the MLA with the best performance. Hence, this study used MLA [10].

The research dataset used was trip data with varied mileage, and driving behavior that was collected randomly. The dataset was obtained from online taxi transaction data provided by one of the technology companies operating in Indonesia that offers transportation service via online order. The application provider obtaining the data by randomly collecting sensor data recordings on the driver’s smartphone. Meanwhile, smartphone sensor data, including accelerometer sensor data, gyroscope sensors, and GPS sensors, were already available. In this study, data were collected during individual trips, covering the entire trips from the pick-up to the drop-off locations. The recorded data included crucial details such as the distance traveled, the specific route taken, and the total duration of each trip. All this information was consolidated and associated with a unique identification number (ID) that represents each booking. It is important to emphasize that the data collection process occurs in real-time and follows a random sampling approach. As a result, the data size for each booking may significantly vary due to the diverse characteristics of different trips. Given the substantial variations in data size across bookings, it becomes essential to conduct data analysis using statistical methods [11]. Statistical analysis enables researchers to gain meaningful insights from the dataset and uncover patterns, trends, and important correlations among the recorded variables. By employing appropriate statistical techniques, this study aims to make informed interpretations and draw valid conclusions from the collected data, contributing to a more comprehensive understanding of the driving behavior and characteristics.

After the initial analysis of the dataset, it was found that the available datasets were imbalanced and the available features were limited, resulting in the RF classifier to yield a low sensitivity value. The existing data problems, namely the imbalanced dataset, in some cases, had significantly affected the accuracy of the data. Nevertheless, some classification algorithms, such as RF, possess characteristics suitable to overcome this imbalanced dataset [12]. Then, a modeling simulation was conducted to test the reliability of the RF classification in this study. If the accuracy results are unsatisfactory, several preprocessing techniques will be implemented first to address the imbalance problem. Several studies have conducted experiments on imbalanced dataset, and random sampling has been proposed as a solution. Some experimental results were successful, while some were not.

As previously mentioned, imbalanced data and limited availability of dataset features contributed to low sensitivity values and unstable data accuracies during the testing. In addition to these problems, numerous errors were still found during the detection of trip data. Therefore, feature extraction was also carried out in addition to using the resampling method. The purpose of feature extraction is to combine several features into a feature that is expected to represent all the data characteristics [13]. Hence, researchers applied resampling and feature extraction methods to overcome the problems found in the database.

The analysis was carried out by comparing the results of the classification of driving behavior using the original dataset and datasets that had passed data preprocessing, starting from cleaning data that had null values, normalizing data values, and extracting features. Then, the next step was to apply the resampling technique. The results of the preprocessing dataset obtained were then used for the classification process using several frequently used classification algorithms. The results are expected to be the first step in implementing the ITS to overcome the problem of negligent driving. Once the classification model was obtained, it was then implemented en masse on existing applications. For the implementation stage, there is the need to conduct further research on the implementation of the model and direct interconnection, using testing data from real-time data of currently operating online transportation applications. However, the implementation of ITS involves more than just technology; the regulations of the country must also be paid attention to.

Machines are capable of emulating human-like thinking by imitating various human activities, such as decision-making, problem-solving, and learning. This cognitive approach enables machines to simulate human-like thought processes. In the realm of human thinking, there are two primary methods: introspection and experimentation. Introspection involves asking questions and engaging in critical thinking to uncover solutions when confronted with challenges. On the other hand, experimentation entails actively trying out different approaches, methods, or strategies to explore and find solutions. By utilizing both introspection and experimentation, machines and humans can enhance their problem-solving abilities and adapt to changing situations effectively.

Behaving like a human can be interpreted as imitating human approach in dealing with problems. The concept of behaving like a human is the reason behind the Turing test. The Turing test is a test of human’s ability to recognize machines by making humans to communicate with entities (machines) via the teletype. If they fail to distinguish the entity being a human or machine within 5 minutes of interaction, for example, the entity which they are communicating with is deemed an intelligent system and passed the test. To pass this test, the entity must at least possess the capability to recognize voices, understand human language, perform speech synthesis and knowledge representation, respond automatically, comprehend machine learning, perform judgments, and make decisions.

Reasoning is a study that allows computers to make perceptions, provide responses, and address a problem. For artificial intelligence systems to achieve this stage, they must model how humans think and respond under ideal conditions.

Behaving rationally means generating a rational attitude in terms of computational processes. Being rational is an act to achieve goals while still considering conditions and self-understanding. In this case, intelligent agents as computer systems play the role of making the best decisions while still considering the situation. For example, when an intelligent agent plays chess, the it is expected to make the best move to win the match.

Big data is a general term for any collection of extensive and complex data sets that traditional data processing tools cannot analyzed. The greater the significance of the collected data, the greater number of new information can be generated. Big data is a collection of data with a very high volume. In this study, the dataset consisted of approximately 16 million data lines to be processed with approximately 2 GB of data. This quantity of data could be analyzed more efficiently using the traditional way. A data stream must be able to receive and process data at high speed and in real-time. The velocity method was used in this research stage because learning techniques still used batch data. However, the streaming data were directly analyzed for the implementation stage. Variety or variation is the number of data types circulating based on their shape and type. In traditional data, the collected data are generally, in the form of structured and fit data. However, in big data technology, the data obtained are generally unstructured and obtained from various sources. In this study, the data variation was apparent due to the use of several trip sensors, such as a gyroscope, accelerometer, and GPS sensors. These three sensors have different data writing formats.

Data mining is an interdisciplinary field of science that combines machine learning techniques with several unique algorithms for pattern recognition, statistics, databases, and visualization to extract meaningful information or knowledge from a large dataset. Using certain techniques, information generated from this process can be used to predict an outcome. Generally, the data mining stages are related to the knowledge discovery and data mining (KDD) process. KDD is a computer-assisted process of recognizing and analyzing large datasets and extracting useful information and knowledge. One of the stages in the entire KDD process is data mining itself. One of functions of data mining implementation, such as the classification method, is grouping objects based on existing groups. The classification method requires training data that has been labelled or classified, which differs it from the clustering technique. The classification method was used in this study.

ITS synthesizes several technologies, such as positioning, communication, information systems, and electronic control. Concerning ITS supporting technology, GPS usually acts as its positioning technology, while geographic information system (GIS) acts as its information system technology. ITS navigation systems can be classified into four types: autonomous ITS, fleet management ITS, advisory ITS, and inventory ITS. ITS combines human factors (people), roads, and vehicles by utilizing state-of-the-art information technology. ITS aims to apply advanced technology to transportation facilities to make them safer, more efficient, better developed, and more environmentally friendly. The following is the scope of the ITS.

Fleet management manages vehicles from the dispatch center through communication links. In this system, the vehicles are equipped with a positioning system, and generally, they are not equipped with an electronic map system. These vehicles report their position to the control center so that the control center has the convenience to manage the movement of these vehicles. Besides providing instructions regarding directions, the control center is also responsible for providing the information needed by vehicle drivers, such as weather information and traffic conditions. Advisory system combines the positioning and electronic map system aspects of the ITS autonomous system with the communication aspects of the ITS fleet management system architecture. The ITS advisory system is autonomous in the sense that a dispatch center does not control this system. However, at the same time, this system is part of the vehicle fleet that receives services from the traffic information center. In some ITS advisory systems, certain vehicles stand alone as traffic probes, providing other vehicles (not defined by the traffic information center) with the latest information about traffic and weather conditions. Inventory system usually consists of a stand-alone vehicle and is equipped with a digital video camera to collect data (complete with coordinates and time of collection) related to the road which are then used for road inventory, road maintenance, and investigation of traffic disturbance objects. The vehicles used are also equipped with positioning devices, data loggers, and data display in the form of electronic maps. Safe driving assistance system is a very advanced form of ITS. The vehicle has several sensors that can direct the driver to drive safely. Numerous research on soft computing in various sciences has been conducted, one of which is in the field of safety driving. Several data mining methods are often used for classification. This method aims to predict an outcome of some or all variables to predict a class containing two or more values [14]. This research combined several technologies to identify a research gap that could be addressed. Researcher have adopted some theories including the theory of artificial intelligence, big data, data mining, and ITS. Researchers found a gap between the slices of interconnected theories. Therefore, in addition to mapping theory, this research also referred to previous studies.

Research on driving detection study has been conducted [10]. The researchers conducted a quantitative evaluation of the performance of MLA classification, SVM, RF, ANN, and BN to detect how to drive using data from four smartphone sensor features such as (accelerometer, linear acceleration, magnetometer, and gyroscope). Research on human activity recognition (HAR), or dubbed human activities, has also been conducted [15]. In this research, the data used were accelerometer sensor data and sound sensors on smartphones. This research classified a person activity, namely sitting, walking, or running. This research compared the performance of the classifiers used in this study, including multi-layer perceptron, decision tree (DT), and SVM. Researchers also raised the topic of imbalance datasets, where from the data described, there was a comparison of data for sitting (26.4%), running (1.9%), and running normally (45.91%). From this problem, the researcher proposed an oversampling method.

The results indicate that the oversampling method can overcome the problem of imbalanced datasets, with the best classifier in this study being the multi-layer perceptron. With the use of oversampling, the classification performance comparison for the F1 multi-layer perceptron value was around 15% [15].

The ML-based monitoring methods with multi-class classification methods was used to identify modes of transportation (cars, bicycles, buses, walking, and running) [16]. The classification methods used were k-nearest neighbor (KNN), SVM, and RF, by processing smartphone sensor data (accelerometer, gyroscope, and light sensors). The feature extraction process was carried out from the sensor data provided, with the final results being 165 features. The RF classifier method produced the best performance based on the test results. In addition, the characteristics of the RF classifier greatly benefited in processing datasets with many features.

Another research has studied human activities, where the data used was smartphone sensor data based on the two-axis accelerometer sensor (x, y) [17]. This study classified normal and aggressive driving using a fuzzy algorithm. The results showed that the fuzzy classification system could present trip graph data based on the expected classification. The study also stated that the accelerometer sensor was more needed when detecting driving behavior. Hence, it is necessary to test using other sensors, such as the gyroscope sensor [17].

A system with machine learning methods to detect and identify specific types of abnormal driving behavior based on 3-axis gyroscope and 3-axis accelerometer data on smartphone sensors has also been proposed [18]. The study used twenty sampling trips and measured unsafe driving based on six criteria: movement such as weaving, swerving, sideslipping, U-turn with a narrow radius (fast U-turn), turning with a wide radius, and sudden braking. The classification method used was SVM and neural networks. Based on the research results, the gyroscope sensor yielded the best classifier accuracy rate among other sensors in detecting driving behavior.

II. METHODOLOGY

In general, the method used in this research was the data mining methodology. Data mining is defined as extracting information from large amounts of data. It includes various statistical analysis and machine learning techniques to determine data patterns and relationships that may be invisible to the naked eye. The goal of data mining is to find helpful information that can be used to make better business decisions and improve the efficiency of data operations. Data mining can also be used to predict future trends or patterns based on historical data [19]. The use of data processing techniques for accelerometers and gyroscope sensors to detect between safe or negligent driving are not new. There are several studies with similar objectives. However, research in this field employed varied methods or schemes. Since this kind of research adapts to the characteristics of the dataset used, the development of new methods or schemes to detect driving behavior remains open for further research. Based on the results of a review of several previous studies, this has contributed to previous research as followings.

The characteristics of the dataset can be important information for comparison in a study. Research on the same topic using different datasets will result in different results and conclusions. This study used a dataset with different characteristics from previous studies. This study used online transportation trip data. There has never been any research on safety driving using this dataset and, it has the most trip data samples compared to previous research, namely 20 thousand trips with data rows of more than 16 million.

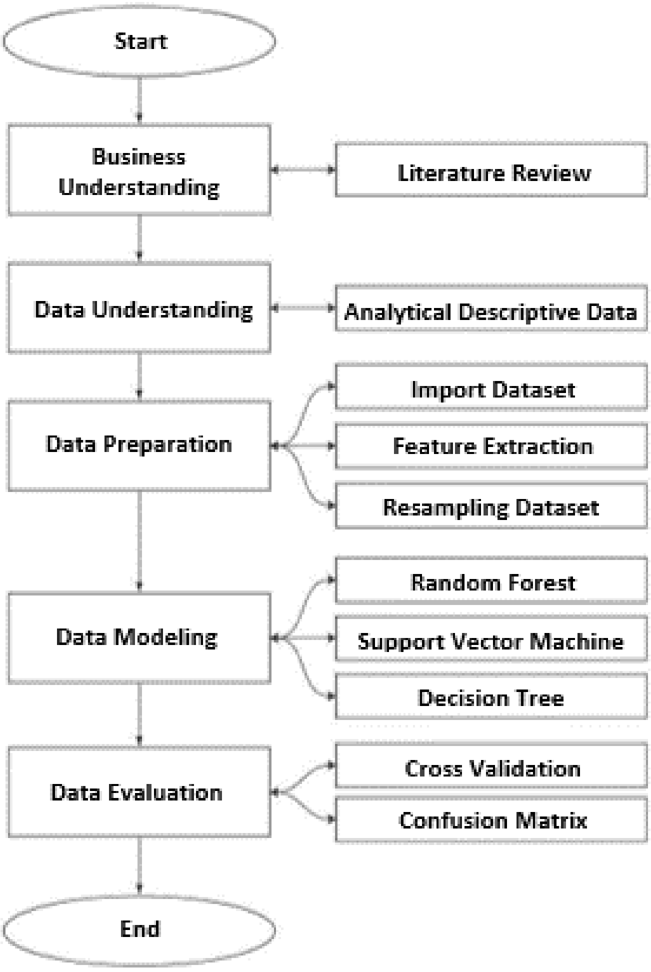

Figure 1. Research flow.

A. DATASET

In this study, the data used were a dataset obtained from mobile phone sensors, which had been exported in .csv format. The dataset contained information from an online transportation transaction system operating in Indonesia. Some crucial aspects obtained from the initial understanding phase include collecting from the online transportation system by capturing trip transactions made using cars with embedded mobile phone sensor data. These sensor data included gyroscope sensor data (x, y, and z axes), accelerometer data (x, y, and z axes), speed data, and time data. The dataset consisted of both training and testing data, and in total, it comprised approximately 16 million rows of data from 20 thousand trips.

B. RESEARCH FLOW

In general, this research flow used existing models in the cross-industry standard process for data mining (CRISP-DM) methodology. CRISP-DM is one of the most well-known and widely used process models in data mining. It provides structure and guidance for managing data mining projects from start to finish, including planning, data collection, data processing, model development, and results evaluation. CRISP-DM is very useful in helping manage data mining projects in an effective and structured manner. In addition, this process model is flexible and can be adapted according to specific project needs [20].

The CRISP-DM process consists of five main stages: business understanding, data understanding, data processing, data modeling, and data evaluation. The used evaluation technique was the measurement of the confusion matrix results, yielding the accuracy value of the constructed model. The accuracy results of the classification model were then compared to its best performance, as illustrated in the Figure 1. In the modeling experiment, two distinct trials were conducted. The objective of the initial trial is to model without implementing any preprocessing techniques, while the second trial involves the application of preprocessing techniques utilized in previous research. Additionally, aside from employing the preprocessing method, several commonly used classification algorithms were compared.

1) BUSINESS UNDERSTANDING

At this stage, a theoretical excavation or literature review was conducted on the processing of smartphone motion sensor data. Although relevant previous research was found during the literature review stage, this research involved an experimental scenario to compare the performance of MLA on trip datasets. These datasets share similar features to this study, such as accelerometer, gyroscope, and GPS data. The characteristics of the dataset used in this research were similar to the dataset in this study, allowing for a comparison between the experimental model used in previous studies and the experimental model in this study.

2) DATA UNDERSTANDING

This stage involved data collection and descriptive analysis of the research data. The dataset consisted of smartphone motion sensor data collected from an online transportation system. The available features included booking ID, time, gyroscope, accelerometer sensors, and speed features. The dataset was retrieved by exporting it into a .csv file from the data provided by the service provider application. The dataset also included labels. Each trip was recorded in real-time, resulting in various data sizes. Therefore, it is necessary to analyze the data using statistical methods.

3) DATA PREPARATION

The primary purpose of this stage was to carry out further data processing because the dataset used was extensive. Dataset available and to be processed had several rows of more than 16 million data rows from 20 thousand transactions. Step by step on the raw dataset, the processed data was the driver’s smartphone sensor data from each transaction. Each transaction possessed different characteristics of time and length of data depending on the distance of the trip route. At the data preparation stage, several things were carried.

The initial step in data preparation wasto import the data from the obtained dataset. As mentioned earlier, there were two types of data used, namely feature data and label data. The label data consisted of two classes, label 0 and label 1, with label 0 representing safe trip and label 1 representing relatively dangerous trip. In this process, the feature dataset was merged with the label dataset, so that each feature data based on the transaction ID had been categorized for each class category. Furthermore, to facilitate reusability and reduce computational effort, the combined result of feature data with label data was stored. In the future, when this data is used, there is no need to merge the datasets again.

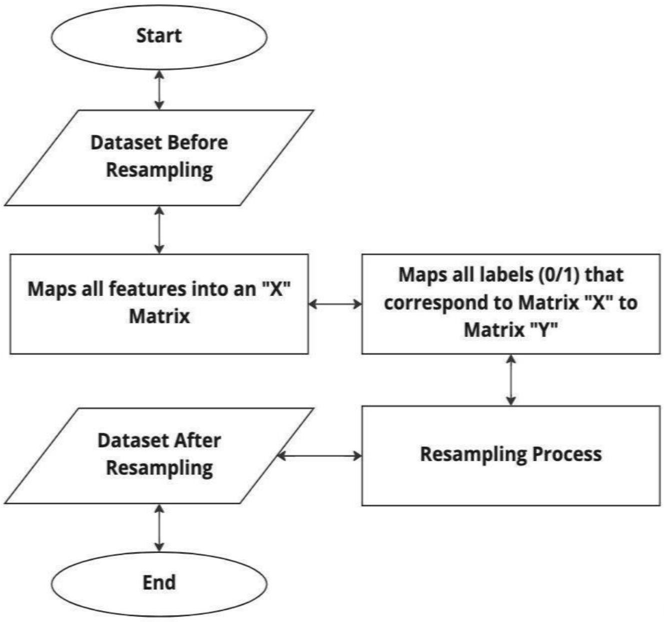
Start
Dataset Before Resampling
Maps all features into an "X" Matrix
Maps all labels (0/1) that correspond to Matrix "X" to Matrix "Y"
Dataset After Resampling
Resampling Process
End

Figure 2. Dataset resampling.

After the stage of merging the feature datasets and labels was complete, the next step was the feature extraction. Based on the available data, the gyro x, y, and z sensors could be combined to create a new feature called “gyroscope.” Similarly, the accelerometer x, y, and z features could be combined to create a new feature called “accelerometer.” Lastly, the time and distance data could be used to create a new feature called “speed.” These new features were derived from the existing features. Before proceeding to the classification stage, feature extraction was performed on each identified trip. The classifier engine utilized these extracted features to differentiate between the two types of labels: safe and dangerous. As previously mentioned, categorizing the trip data into these two labels facilitates data integration during the feature extraction stage. The data, after undergoing the feature extraction process, formed new features that enabled the classifier to quickly identify the appropriate trip label.

Following the feature extraction was dataset resampling. This stage is a proposal proposed to overcome dataset imbalance found in the descriptive analysis at the beginning of the study. From the datasets that had been checked, the imbalance was found in the number of data labels for safe and dangerous trips. The number of data label 0 is greater than that of label 1. This imbalance was a problem for the classifier engine because the dominant amount of safe trip data made it easy for dangerous trip objects to be recognized as safe trips.

Figure 2 presents the flow of applying the dataset resampling used in this study. The method used was the dimension reduction method that was commonly used in ML, namely principal component analysis (PCA). The PCA method was used to map objects in the coordinate plane. With this method, the data distribution can be known. Therefore, it can be concluded using the appropriate resampling method in each dataset. Along with the characteristics of the data distribution on the research dataset from the results of the analysis based on the PCA method, the resampling method of the dataset in this study could be used to examine the effect of imbalance on the dataset and whether it affected the classification results of machine learning.

4) DATA MODELING

After going through the data preprocessing stage, the preprocessing dataset yielded features with a high significance for distinguishing between two classes: safe trip and dangerous trip. These features were used in the proposed classification experiment. Currently, no model can detect safety driving. With this research, the driving characteristics classification model can be a detection model that can detect unsafe trip. Therefore, this stage attempted to reduce the detection error by applying classification techniques. It enhanced the performance and accuracy of driving recognition detection in the experimental method proposed, which included a classification model employing multiple MLA: RF, SVM, and DT. Based on the PCA distribution characteristics, these three classification algorithms were appropriate to be used with datasets containing existing characteristics.

5) DATA EVALUATION

The testing was carried out in this stage. In this study, three experimental schemes were performed, including the followings.

- Testing the RF classification model on a dataset with processing stages.
- Testing the SVM classification model on a dataset with processing stages.
- Testing the DT classification model on a dataset with processing stages.

After conducting three experimental schemes of the classification methods, a validation process was carried out to evaluate the performance of the classification model. This validation process utilized five-fold cross-validation, commonly known as *k*-fold cross-validation. It has been observed in some studies that *k*-fold cross-validation exhibits lower variance compared to a simple training-test data split. This lower variance implies that the data points tend to be closer to the expected value. This characteristic is particularly valuable when dealing with limited data. In *k*-fold cross-validation, the data is divided into *k*-folds, typically using a value of 5. Each fold of the data serves as the test data once, while the remaining folds are used as training data. This process is repeated *k* times, with each fold serving as the test data exactly once. The resulting confusion matrix provides the final outcome of this validation process. By comparing the confusion matrices from multiple experiments, it is possible to determine which classification model performs best in classifying smartphone motion sensor data and identifying driving characteristics.

III. RESULT AND DISCUSSION

The data processed in this study were trip detection sensor data such as gyroscope, accelerometer, and GPS sensors installed on the driver's smartphone. The dataset was obtained from the results of trip records of transportation service providers operating in Indonesia. Before entering the data processing stage, the exploratory data analysis was conducted. This stage was intended to recognize and learn more about the data's characteristics. There must be no noise or missing values in the data to be processed. Thus, the processing could be conducted with the appropriate approach to obtain the desired results. The results of the exploratory data analysis provided a clear representation. Trip data had a total of fourteen features. Further information regarding this data can be seen in Table I.

TABLE I
FEATURE DATASET

Data	Descriptions
Accuracy	Accuracy measurement value on the GPS module
Bearing	The measurement value of the number of revolutions of the digital compass
acc_x	The x-axis digital acceleration measurement value
acc_y	The y-axis digital acceleration measurement value
acc_z	The x-axis digital acceleration measurement value
Vector_accelerometer	Feature extraction from acc x,y and z vectors
gyroscope_x	The measurement value of the x-axis digital gyroscope
gyroscope_y	The measurement value of the y-axis digital gyroscope
gyroscope_z	The measurement value of the z-axis digital gyroscope
Vector_gyro	Extraction features from the Gyroscope x,y and z vectors
second	Data time measurement value in seconds
Speed	The value of moving speed is measured in meters
Distance	Speed and time extraction feature to get distanceSpeed and time extraction feature to get distance
Label	The feature dataset label is already available by the data owner. The labeling method was carried out by collecting sampling based on an assessment of customer satisfaction in one on-going transaction from the pick-up point to the drop-off point with passenger review parameters. Driving behavior was classified manually on each trip with assumptions class “0” as safe and Class “1” as dangerous.

1) DATA MODELING

The performance test parameters that were measured in this study are as followings. Accuracy is a matrix that measures the number of the model's predictions that are correct from the total amount of predicted data. This metric is generally used when the target class has a balanced distribution. Precision measures how many optimistic predictions are correct from the total positive predictions. This metric ensures that the model does not classify harmful data as positive. Recall counts how many positive prediction results are valid from the complete positive data. This metric is helpful to ensure that the model can identify as much positive data as possible. F1-score is a harmonic average between precision and recall. This metric is used to account for both accuracies and recall simultaneously. The model validation stage is carried out to evaluate the performance of statistical models or machine learning on data that are not used in the model training process. Model validation is essential to ensure that the model generalizes well to unknown data. In this study, the split validation and cross-validation approach followed the results of model performance testing.

2) RESULTS OF SPLIT VALIDATION TEST

This method separates data into two parts: training data and data validation (or testing). The model was trained on training

TABLE II
RESULTS OF SPLIT VALIDATION TEST

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	76.00%	76.00%	100.00%	87.00%
RF	91.00%	92.00%	97.00%	94.00%
DT	88.00%	92.00%	92.00%	92.00%

TABLE III
RESULTS OF CROSS-VALIDATION TEST

Accuracy Model Results			
K-Folds	SVM	RF	DT
1	76.61%	91.30%	87.42%
2	76.31%	91.39%	87.40%
3	76.47%	91.45%	88.01%
4	76.70%	91.02%	87.93%
5	76.39%	90.94%	87.57%
Mean	76.50%	91.22%	87.67%

data and evaluated on data validation. The separation ratio was 80:20. The following are the results of performance testing using the split validation method on the SVM, RF, and DT.

Based on the results of model testing using split-validation in Table II, the accuracy, precision, recall, and F1-score values were obtained based on the comparison of performance values of all the parameters of the RF test, which have the best performance. In the SVM model, the test results found that the accuracy value was 76% and the recall value was 100%, meaning that the SVM model has good performance in identifying positive classes (true positives). Yet, this model resulted in many false negatives, indicating that the model cannot remember all safe class trips as it should be. It can occur when the model prefers to classify trips as dangerous (false negative) to maximize the accuracy of the training data.

3) RESULTS OF CROSS-VALIDATION TEST

This method splits data into parts (usually 5 or 10) called folds. The model was trained on *k*-fold data (*k*-1 fold as training and one-fold as validation) and was evaluated on *k*-folds that were not used as training/validation. It was done *k* times so that each piece of data was used as a validation set. Finally, the average value of the *k* evaluation results was taken as the final result. The cross-validation parameter used was 5. The results of the cross-validation are presented in Table III.

4) RESULTS OF STANDARD DEVIATION TEST

The standard deviation is a measure of the data distribution, which indicates how far the data are from the average value. In a classification model, the standard deviation can provide information about how variable or spread the data are in each class. For example, if the standard deviation of features in a particular type is low, the values of these features tend to approach each other and are less varied. Conversely, if the standard deviation of the elements in a particular class is high, the values of these features are more varied. In making a classification model, information on the variability of features in each category can help select and evaluate the suitable model. For example, if there are significant differences in the data distribution between classes, then the built model may consider using different methods for each of these classes.

TABLE IV
RESULTS OF STANDARD DEVIATION TEST

Classifier	Standard Deviation Results
SVM	0.0014
RF	0.0020
DT	0.0026

TABLE V
RESULTS OF THE MOST INFLUENTIAL FEATURE TEST

Features	Feature Importance	Feature Correlation
Speed	0.14580	0.1175
acc_z	0.10225	0.0833
Accuracy	0.07678	0.0649
acc_y	0.06993	0.0509
second	0.06780	0.0918
Bearing	0.06721	0.0032
Distance	0.05161	0.1188
acc_x	0.03925	0.0300
Vector_Gyro	0.03147	0.0611
Vector_Acc	0.02027	0.0280
gyroscope_z	0.01681	0.0036
gyroscope_x	0.01600	0.0179
gyroscope_y	0.01563	0.0032

In this study, researcher obtained a comparison of the standard deviation in the training dataset, which was carried out using the SVM, RF, and DT using standard parameter tuning and the number of *k* = 5. Table IV presents a comparison of the results obtained.

Based on the calculation, the standard deviation value obtained were 0.0014 for SVM, 0.0020 for RF, and 0.0026 for DT. The results showed that overall, the three models produced good values with parameters below 0.1. There is no universal “good” standard deviation value for cross-validation results. This value can vary depending on many factors, such as dataset size, number of cross-validation folds, and model complexity; however, the smaller the standard deviation, the better the model performance. A standard deviation of less than 0.1 indicates that the model has consistent performance across fold cross-validation. In contrast, a standard deviation of greater than 0.1 indicates that the model has a more significant variation in performance across fold cross-validation. However, it is essential to remember that the standard deviation should be assessed with the accuracy score and other evaluation metrics to get a more holistic picture of model performance. In addition, model parameter tuning needs to be done carefully because it can affect the standard deviation value and overall model performance.

5) MOST INFLUENTIAL FEATURES

From modeling results, it can be seen the most influential features in classifying driving characteristics. Feature assessment selects a dataset’s most relevant and significant subset of elements for analysis or machine learning model building. There are several reasons why feature selection is so important. If the dataset has too many irrelevant features, the machine learning model can become too complex and overfitting. Feature selection helps reduce model complexity and prevent overfitting. Machine learning models used in datasets with relevant and significant features usually perform better than those with irrelevant features. Datasets with fewer features will require less time and cost to process and analyze. Datasets with relevant and significant features are also easier to interpret than those with irrelevant ones, allowing for easier and more accurate use of analytical results. Overall, feature selection is critical to producing machine learning models that are accurate and efficient for processing and analyzing datasets. Some of the methods used to find out the essential features in the research dataset are as followings. The first method, dubbed feature selection based on importance, machine learning models to evaluate the importance of each feature in the dataset and selects the most relevant components to the target variable. Then, the second method (or feature correlation) performs correlation analysis between each pair of elements in the dataset and likes features with a low or insignificant correlation. From the two methods above, the following is the result of the calculation.

Table V show that the feature with the most influence on the class is the Speed feature. It indicates that speed is the main factor in affecting safety driving characteristics.

IV. CONCLUSION

Based on the results of the data processing experiments, the dataset used in this research is extensive. After being combined with the label data, it turns out that not all trips were labeled, so the preprocessing stage was carried out by removing data that did not have a label. Then, it was also found that dataset wan imbalanced, with a composition of 20:80. The dataset tended to contain safe trip data labels. With these characteristics, the classification method produced a good performance value without carrying out the resampling process. The final results of this study showed that the random forest had the best performance results, with an accuracy of 91.22%. In addition, this study generated which features or data variables affected driving characteristics. The results showed that speed was the main factor affecting driving characteristics. Besides speed, driving behavior was the second factor that can capture operating characteristics safely or seriously. The conclusion of this study can be used to determine the influence of the imbalanced dataset and whether that dataset affects the performance of machine learning classification. Then, it is expected to know the driving characteristics by processing the motion sensor data on a smartphone. Finally, the model created is expected to be implemented to detect types of safe and dangerous trips. This research will be the initial stage of implementing the Intelligent Transport System (ITS) for driving safety.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest in this paper and during the research conducted.

AUTHOR CONTRIBUTION

Conceptualization, Lisa Dinda Yunita, Ema Utami, and Ainul Yaqin; methodology, Lisa Dinda Yunita and Ema Utami; model analysis, Lisa Dinda Yunita and Ainul Yaqin; validation, Lisa Dinda Yunita, Ema Utami, and Ainul Yaqin; formal analysis, Lisa Dinda Yunita; model validation, Lisa Dinda Yunita; data curation, Lisa Dinda Yunita; writing—original drafting, Lisa Dinda Yunita; writing—reviewing and editing, Lisa Dinda Yunita, Ema Utami, Ainul Yaqin; visualization, Lisa Dinda Yunita.

REFERENCES

[1] World Health Organization, “Global Status Report on Road Safety 2018,” 2018, [Online], https://www.who.int/publications/i/item/978924156568

[2] A. Pirayre, P. Michel, S.S. Rodriguez, and A. Chasse, “Driving Behavior Identification and Real-World Fuel Consumption Estimation with Crowdsensing Data,” *IEEE Trans. Intell. Transp. Syst.*, Vol. 23, No. 10, pp. 18378–18391, Oct. 2022, doi: 10.1109/TITS.2022.3169534.

[3] P. Kukuca and M. Chlebovec, “Vehicle Location System,” *2006 Int. Conf. Appl. Electron.*, 2006, pp. 101–104, doi: 10.1109/AE.2006.4382974.

[4] J.F. McLellan, M.A. Abousalem, dan T.R. Porter, “Quality Control in DGPS Separation Vector Systems,” *Proc. 1994 IEEE Position Locat., Navig. Symp. - PLANS'94*, 1994, pp. 726–732, doi: 10.1109/PLANS.1994.303410.

[5] L. Zhu *et al.*, “Big Data Analytics in Intelligent Transportation Systems: A Survey,” *IEEE Trans. Intell. Transp. Syst.*, Vol. 20, No. 1, pp. 383–398, Jan. 2019, doi: 10.1109/TITS.2018.2815678.

[6] M. Adnane, B.-H. Nguyễn, A. Khoumsi, and J.P.F. Trovão, “Driving Mode Predictor-Based Real-Time Energy Management for Dual-Source Electric Vehicle,” *IEEE Trans. Transp. Electrific.*, Vol. 7, No. 3, pp. 1173–1185, Sep. 2021, doi: 10.1109/TTE.2021.3059545.

[7] A. Fang, C. Qiu, L. Zhao, and Y. Jin, “Driver Risk Assessment Using Traffic Violation and Accident Data by Machine Learning Approaches,” *2018 3rd IEEE Int. Conf. Intell. Transp. Eng. (ICITE)*, 2018, pp. 291–295, doi: 10.1109/ICITE.2018.8492665.

[8] S. Agarwal, “Data Mining: Data Mining Concepts and Techniques,” *2013 Int. Conf. Mach. Intell., Res. Adv.*, 2013, pp. 203–207, doi: 10.1109/ICMIRA.2013.45.

[9] J.A. Talingdan, “Performance Comparison of Different Classification Algorithms for Household Poverty Classification,” *2019 4th Int. Conf. Inf. Syst. Eng. (ICISE)*, 2019, pp. 11–15, doi: 10.1109/ICISE.2019.00010.

[10] G. Castignani, R. Frank, and T. Engel, “Driver Behavior Profiling Using Smartphones,” *16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC 2013)*, 2013, pp. 552–557, doi: 10.1109/ITSC.2013.6728289.

[11] T.-Y. Lee and H.-W. Shen, “Efficient Local Statistical Analysis via Integral Histograms with Discrete Wavelet Transform,” *IEEE Trans. Vis., Comput. Graph.*, Vol. 19, No. 12, pp. 2693–2702, Dec. 2013, doi: 10.1109/TVCG.2013.152.

[12] L. Hakim and S. Rochimah, “Oversampling Imbalance Data: Case Study on Functional and Non Functional Requirement,” *2018 Elect. Power Electron. Commun. Controls, Inform. Seminar (EECCIS)*, 2018, pp. 315–319, doi: 10.1109/EECCIS.2018.8692986.

[13] F.P. Shah and V. Patel, “A Review on Feature Selection and Feature Extraction for Text Classification,” *2016 Int. Conf. Wirel. Commun. Signal Process., Netw. (WiSPNET)*, 2016, pp. 2264–2268, doi: 10.1109/WiSPNET.2016.7566545.

[14] Y. Xiao, Y. Liu, Y. Deng, and H. Li, “Enhancing Multi-Class Classification in One-Versus-One Strategy: A Type of Base Classifier Modification and Weighted Voting Mechanism,” *2021 Int. Conf. Commun. Inf. Syst., Comput. Eng. (CISCE)*, 2021, pp. 303–307, doi: 10.1109/CISCE52179.2021.9445948.

[15] K.T. Nguyen, F. Portet, dan C. Garbay, “Dealing with Imbalanced Data Sets for Human Activity Recognition Using Mobile Phone Sensors,” *3rd Int. Workshop Smart Sens. Syst.*, 2018, pp. 1–10.

[16] A. Jahangiri dan 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, doi: 10.1109/TITS.2015.2405759.

[17] A. Aljaafreh, N. Alshabatat, and M.S.N. Al-Din, “Driving Style Recognition Using Fuzzy Logic,” *2012 IEEE Int. Conf. Veh. Electron., Safety (ICVES 2012)*, 2012, pp. 460–463, doi: 10.1109/ICVES.2012.6294318.

[18] J. Yu *et al.*, “Fine-Grained Abnormal Driving Behaviors Detection and Identification with Smartphones,” *IEEE Trans. Mobile Comput.*, Vol. 16, No. 8, pp. 2198–2212, Aug. 2017, doi: 10.1109/TMC.2016.2618873.

[19] S.M. Dol and P.M. Jawandhiya, “Use of Data mining Tools in Educational Data Mining,” *2022 Fifth Int. Conf. Comput. Intell.,*

Commun. Technol. (CCICT), 2022, pp. 380–387, doi: 10.1109/CCiCT56684.2022.00075.

[20] F. Schäfer, C. Zeiselmair, J. Becker, and H. Otten, “Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes,” *2018 IEEE Int. Conf. Technol. Manage. Oper., Decis. (ICTMOD)*, 2018, pp. 190–195, doi: 10.1109/ITMC.2018.8691266.