A Novel Compound Feature based Driver Identification

Md. Abbas Ali Khan[1], Mohammad Hanif Ali[2], AKM Fazlul Haque[3], Md. Iktidar Islam[4], Mohammad Monirul Islam[5]
Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh[1, 3, 4, 5]
Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh[1, 2]

***Abstract*—In today's world, it is time to identify the driver through technology. At present, it is possible to find out the driving style of the drivers from every car through controller area network (CAN-BUS) sensor data which was not possible through the conventional car. Many researchers did their work and their main purpose was to find out the driver driving style from end-to-end analysis of CAN-BUS sensor data. So, it is potential to identify each driver individually based on the driver's driving style. We propose a novel compound feature-based driver identification to reduce the number of input attributes based on some mathematical operation. Now, the role of machine learning in the field of any type of data analysis is incomparable and significant. The state-of-the-art algorithms have been applied in different fields. Occasionally these are tested in a similar domain. As a result, we have used some prominent algorithms of machine learning, which show different results in the field of aspiration of the model. The other goal of this study is to compare the conspicuous classification algorithms in the index of performance metrics in driver behavior identification. Hence, we compare the performance of SVM, Naïve Bayes, Logistic Regression, k-NN, Random Forest, Decision tree, Gradient boosting.**

***Keywords*—Compound feature; driver behavior identification; engine speed; fuel consumption; vehicle**

I. Introduction

Every driver has their driving style; therefore, the driver can be identified according to exploration through the driving pattern analysis. It is to be considered as a fingerprint of the driver's manner like acceleration, speed, and braking habits that vary from driver to driver. Driver fingerprinting could lead to important privacy compromises [1]. Today we cannot consider just a vehicle as a modern car, as it is a fully decorated smart device with various functions like multimedia, security system, and different sensors [2]. The sensors were very simple because the driver was informed regarding the features of the engine and the amount of fuel through the magnetoelectric and light display devices [2]. Using state-of-the-art technology in real-time all the microcomputers are communicated with each other through CAN-BUS (Controller Area Network) [3]. To make a car more efficient a good number of technologies are used in the modern engine. To improve engine performance direct injection technology was introduced in the modern car [4]. According to a survey, the researcher predicted that the number of sales of connected cars will reach 76.3 million in the next 2023 [5]. Through state-of-the-art technology, modern engines use less fuel and besides get more power [6]. Most of the cars have partnered with other components which are highly technology-based, such as traffic lights, garage doors, and services [7]. Not only the driving style there is a discount policy on insurance services but also real-time monitoring, maintenance, pathfinding, driving style development, and also consumption of fuel [8]. Vulnerabilities of connected cars will increase the auto-theft which is one of the threats [9]. Top-of-the-range vehicles are targeted by thieves who simply drive off after bypassing security devices by hacking onboard computers [10]. Penny [11] introduced a man-in-the-middle attack or relay attack, to do this radio signals are passed between two devices. Pekaric I et al. (2021) [12] described other attacks such as GPS spoofing and message injection attacks. BMW Connected Drive [13] seamlessly integrates mobile devices, smart home technology, and vehicle's intelligent interfaces into a complete driver's environment. Even though in 2021 they introduced a remote door unlock system through a signal to the driver's door to unlock [13]. CAN-BUS is likely a nervous system used to allow configuration, data logging, and communication among electronic control units (ECU) e.g., ECU is like a part of the body and interconnected through CAN, by which information sensed by one part can be shared with another [14]. Up to 70 ECUs have a modern car e.g., the engine control unit, airbags, audio system, acceleration, fuel unit, etc. [15]. Normally, multi-sensor data is made up of in vehicle's CAN data. The in-vehicle CAN data such as steering wheel, vehicle speed, engine speed, amount of fuel, etc. Several researchers previously proposed a driver identification method based on in-vehicle CAN-BUS data. But direct connectivity is difficult to get data, so onboard diagnostics (ODB-II) is used. (OBD-II, ISO 15765) is a self-diagnostic and reporting capability that e.g., mechanics use to identify car issues, OBD-II specifies diagnostic trouble codes (DTCs) and real-time data (e.g., speed, revolution per minute RPM), which can be recorded via OBD-II loggers from CAN-BUS. Many authors described the problem of CAN-BUS data for identifying the driver [9], [16], [17].

Since this is a big dataset and there are 51 features with 10 labels. Moreover, for analyzing the whole dataset we need more time. To reduce the time complexity we have explained compound feature selection process. In this paper, our objective is to identify driver behavior through telemetric data using machine learning algorithms. We analyze the data in terms of training, testing, and validation to get model accuracy that helps us with driver identification.

There are some significant commercial and personal values of driver identification for insurance, rental companies, personal and tripping regulations as well. The goal of this work is to monitor and maintain unauthorized access.

The paper is organized as follows. We have presented an introduction and literature review Sections I and II as well. The architecture of the proposed system has described in Section III. The discussion of methodology and compound features have existed according to Sections IV and V. Before the conclusion part of Section VII, the results and discussion are mentioned in Section VI.

II. Literature Review

The author [18] uses telemetric data to investigate driver identification and identification accuracy decreases by 15% compared to the method. They use the role of non-public parameters in identifying the driver. Previous work had been done by using car driving simulated [18] data. Investigated the driver's behavior when he follows another car. The features mentioned below are used to observe such as accelerator pedal, car speed, brake pedal, and distance to the next car. Gaussian Mixture Model (GMM) is used to achieve 81% accuracy with 12 drivers and 73% of 30 drivers [18]. Another author [19] analyzes overtaking style for each driver, uses accelerator, and steering data and the accuracy is 85% for about 20 drivers through the Hidden Markov Model (HMM). Some other authors used smartphones to capture driver data. Sensors in the smartphone are- GPS, accelerometer, magnetometer, and gyroscope [20-22]. This data is used for driver profiling and other tasks. An author used an inertial sensor and algorithm was SVM, and k-means methods and got 60% accuracy between two drivers.

In another research, the authors [8, 16], [23-24] acquire data from in-vehicle CAN-BUS via OBD-II. The author [23] uses in-vehicle CAN-BUS sensor data and the accuracy was 99% among 15 drivers. He uses SVM, Random Forest, Naive Bayes, and k-NN methods. Another researcher [9] got 99% accuracy from 51 features of 10 drivers and used Decision Tree, k-NN, Random Forest, and Multilayer Perceptron (MPL). Choi et al. [24] find out the driving detection and driver recognition using both GMM and HMM methods, which are used for analyzing vehicle CAN-BUS data. Kedar-Dongakar et al. [25] recognized the driver classification based on the energy optimization of a vehicle. Based on driving style three types of drivers are classified as aggressive, moderate, and conservative. The author considers the following features for his research work such as vehicle speed, acceleration, torque, acceleration pedal, steering wheel angle, and brake pedal pressure.

Several researches have been going on neural network and deep learning algorithms for a few years back and draw a good impact on driver behavior identification works. Xun et al. [26] introduced Convolutional Neural Network (CNN) and got 99% accuracy for 10 drivers. For driver identification, another paper uses noise-free data and as an algorithm-LSTM-Recurrent Neural Network is used, where it has high accuracy with salient advantages [27].

After reviewing, the difference among this work and the done before.

- For collecting dataset uses specific model's vehicle.
- Specific dataset with different features
- Uses CAN-BUS and public OBD-II.

III. Architecture of the Proposed System

The purpose of the proposed system is to identify the actual driver. After capturing the raw data through OBD-II sent to the cloud is shown in Fig. 1. Once completed the engineering knowledge-based mathematical operation like adding, multiplication, and logging of two features then send for prediction based on model accuracy result.

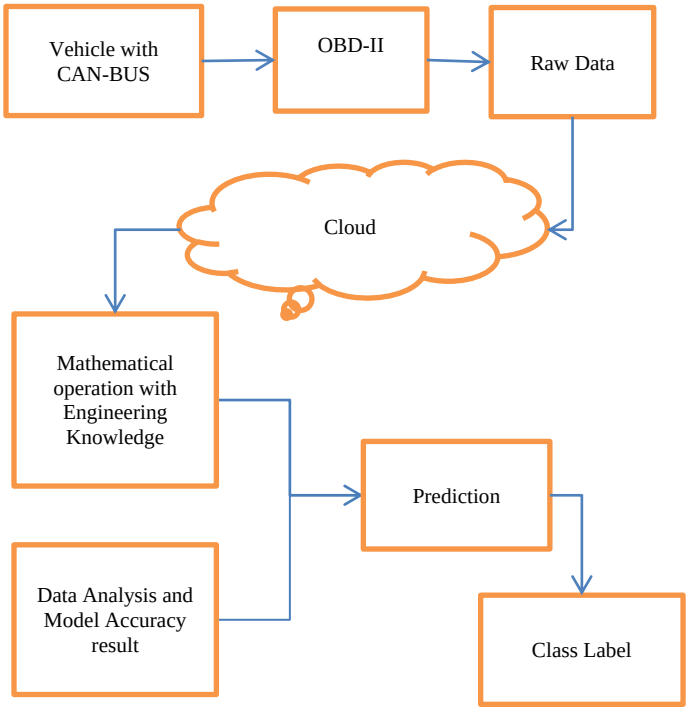

Fig. 1. Architecture of the proposed system

IV. Methodology

A. Approaches of Methodology

The steps of the proposed system are shown in Fig. 2. Before the classifier's dataset preparation, data preprocessing, feature selection, normalization, and other activities have occurred. Several classifiers are used for comparative analysis with state-of-the-art classifiers to find the best model.

In this connection, we need data on the trips for driver identification. Our model is considered an Ocslab driving dataset [28]. This data is used for driver indemnification and personalization based on pattern analysis. KIA motors corporation vehicles in South Korea were performed to collect the data and the experiment has been done since July 28, 2015 [28]. A total of 10 drivers labeled "A" to "J" are included in the trips and cover 23 km length, completing two round trips from 8.00 PM to 11.00 PM [28]. Three types (such as city roads, freeways, and parking lots) of the road are there with

their characteristics. There are a total of 94,401 records with 51 dimensions (51 features) [28]. Not all data is possible to get because there are some limitations of OBD-II identifiers and sensors such as it cannot provide body control status or airbag status even wheel angle rotation status. OBD-II has a limited set of identifiers [29] provided by the manufacturer.

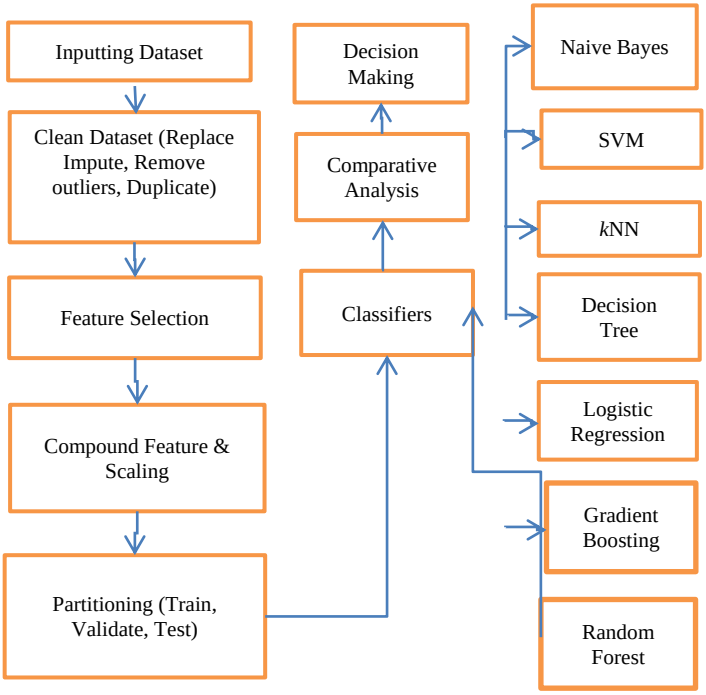

Fig. 2. Step by step process of the work

Among 51 features we have considered two prominent features with around 13,000 samples named Fuel Consumption (FC) and Engine Speed (ES) to make a compound feature. Transform the collected data to our classification model for analysis we follow- feature selection, making a compound feature, scaling, and data processing through state-of-the-art techniques [34]. The sample data are viewed in Eq. (1).

$$X = \begin{bmatrix} X_1^1 & X_2^1 & X_d^1 \\ X_1^2 & X_2^2 & X_d^2 \\ \vdots & \vdots & \vdots \\ X_1^N & X_2^N & X_d^N \end{bmatrix} \quad (1)$$

Where d columns correspond to d variable and N rows correspond to N instances. Fig. 3 shows the histogram of the compound dataset.

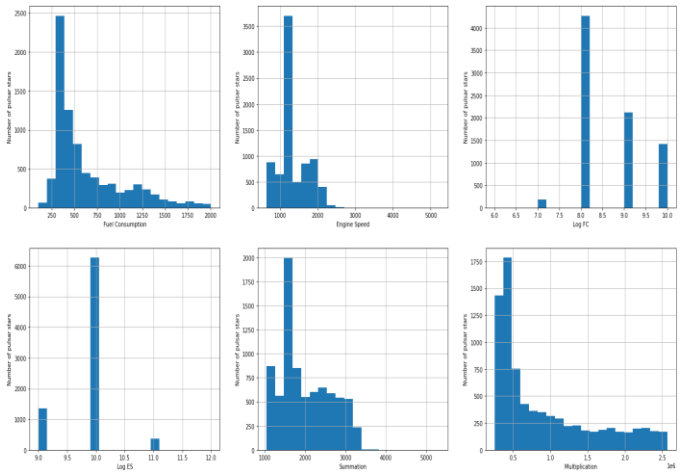

Fig. 3. Histogram of the dataset

To improve the dataset several statistical and logical tools are used to overcome the problem. Significantly different data points are removed by outliers [30].

B. Data Visualization and Relationship in a Dataset

a) Heat-map: A heat map is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions [31-32]. The color variation may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space [38]. We consider a correlogram a clustered heat map that has the same trait for each axis to display how the traits in the set of traits interact with each other. The correlogram is a triangle instead of a square because the combination of A-B is the same as the combination of B-A and so does not need to be expressed twice. Fig. 4 is shown the feature correlation of the compound dataset. The value of correlation can take any value from -1 to 1, based on that-

- Features such as Fuel Consumption, Log FC, Sum, and Mul are having strong positive correlations.
- Fuel consumption and engine speed, engine speed, and Log FC, Log FC, and Log Es have a weak positive correlation.
- Class has a strong negative and weak positive correlation with every feature.

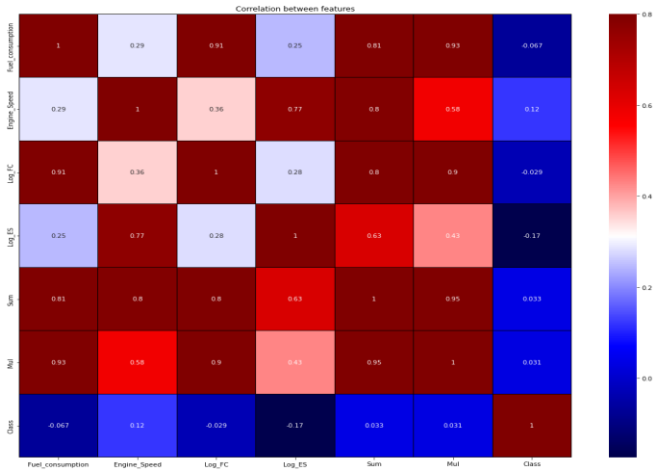

Fig. 4. Correlation among features

C. Pair Plot

A pair plot is a pairwise relationship in a dataset [36]. The pair plot function creates a grid of Axes such that each variable in data will be shared on the y-axis across a single row and the x-axis across a single column [39]. By plotting pair plots, it visualized that most of the classes of each feature overlap with each other. Fig. 5 shows the pair plot of the compound dataset. After analyzing the scatter plot a statistically significant result is shown that non-linear classification models such as k-NN, Decision tree, Random Forest classifier, Gradient boosting classifier, etc. perform better than linear classification models such as logistic regression.

D. Standard Scaler

Standard Scaler comes into play when the characteristics of the input dataset differ greatly between their ranges, or simply when they are measured in different units of measure [37]. Standard Scaler - standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.

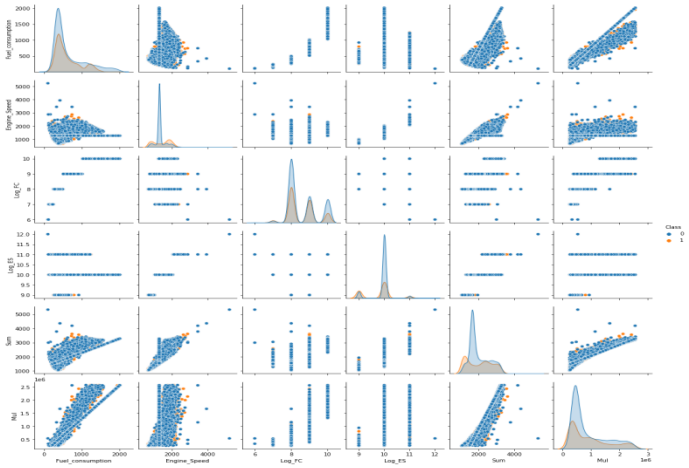

Fig. 5. Pair plot of the dataset

As we see, different scales of data exist in the dataset. Normalization is essential for some machine learning algorithms like k-NN (k-Nearest Neighbor) and SVM [29]. The normalization formulas for integrating data scales are shown in Eq. 2.

$$Z = {(x-\mu)}/{\sigma} \quad (2)$$

Here, μ = mean and σ= standard deviation. To evaluate the generalization of the model we have considered K fold cross-validation for low bias and a modest variance [33]. We have used five folds for training and testing the dataset and obtained the mean performance.

In the case of classification application, we have considered supervised machine learning classifiers for performing the metrics named *k*-NN, SVM, Logistic Regression, Decision tree, Random Forest, and Gradient boosting. The *k*-NN (*k*-Nearest Neighbor) is an instance-based traditional machine learning algorithm. Both classification and regression cases *k*-NN can be used and select the number of neighbors through distance calculation of the query points. Eq. 3 is used to calculate the Euclidean distance between two points.

$$D = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (3)$$

SVM stands for Support Vector Machine, used for classification and regression problems. The goal is to find a hyperplane in an N-dimensional space and separately classify the query data point. There is a decision boundary called hyperplane that is used to differentiate the classes. It also creates a margin separator with the nearest observations and it performs better if maximizes the margin. The equation represents the loss function that indicates maximize the margin

$$C(x,y), Where\ Y = f(x) = \begin{cases} 0, if\ y * f(x) \geq 1 \\ 1 - y * f(x), else\ 1 \end{cases} \quad (4)$$

Logistic Regression predicts whether something is true or false. Instead of fitting a line to the data, it fits an "S" shaped "logistic function" and the curve goes from 0 to 1. The following equation is used to calculate the function, also called the sigmoid function.

$$S(x) = \frac{1}{1+e^{-x}} \quad (5)$$

Naïve Bayes is a classifier based on Bayes' theorem. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The naïve Bayes model is easy to build a large dataset and outperforms sophisticated classification methods. The way Naïve Bayes is used to calculating the posterior probability shows in the Eq. 6.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (6)$$

To measure the performance, we have used four indicators named accuracy, precision, F-Measure, and Recall. The computation and the evaluation performance of the classifiers have occurred through the confusion metric. Once the model is generated then the classifier is tested by using a test dataset to check the model's accuracy. Precision indicates how close or dispersed the measurement is to each other. It measures the number of correct positive predictors made. A recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

The number of *FP*'s, *FN*'s, *TP*'s, and *TN*'s cannot be calculated directly from this matrix. The values of *FP*'s, *FN*'s, *TP*'s, and *TN*'s for class $\boldsymbol{i}$ ($1 \le i \le n$) are determined as per [35].

The final confusion matrix, which has dimension 2 × 2, comprises the average values of the n confusion matrices for all classes. For a binary, i.e. two-class problem, a confusion matrix gives the number of false positives (*FP*s), false negatives (*FN*s), true positives (*TP*s), and true negatives (*TN*s). From this confusion matrix, accuracy, precision, recall, and *F*-1 score are calculated in the following way:

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (8)$$

$$F_1 - score = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \quad (9)$$

$$Accuracy = \frac{TP+TN}{(TP+FN)+(FN+TN)} \times 100\% \quad (10)$$

Fig. 6 shows the time series of a dataset with a window based on RPM. We have considered time t_0 with the window size of w_1 and t_{0+1} time for the next window and so on. Fig. 9 shows the accuracy of the several models based on time and window size.

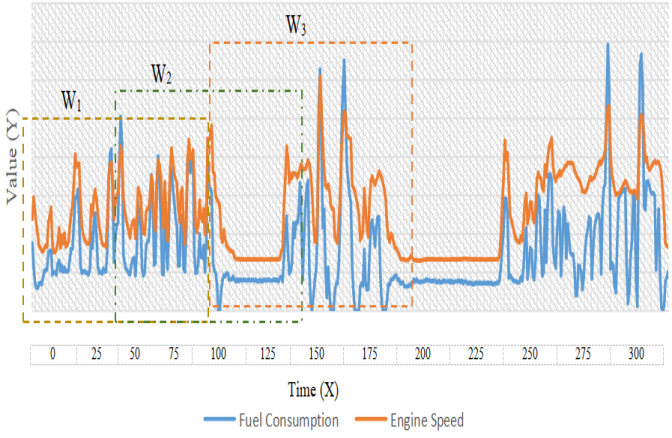

Fig. 6. Time series of dataset with window

V. Compound Feature

To make the feature compound we consider a general mathematical operation. We have selected two important features among the 51 features from the Ocslab dataset. Once we have applied mathematical operations to the selected feature to achieve the compound feature. The compound feature processing method is stated in Fig. 7.

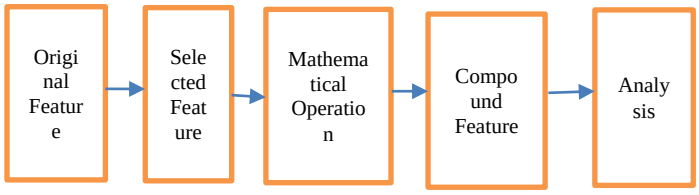

Fig. 7. Preparation of compound feature

From the mathematical operation, we have considered addition, multiplication, and binary logarithm to make the feature compound. Eq. (11), (12), and (13) describe the mathematical operation. Here, k > 0, $k \in \mathbb{N}$ and a_k, b_k are two features with the index of *k*.

Since the value of the features are numerical so we add two features to make a single (Compound) feature is shown in Eq. (11). Eq. (12) is shown for multiplication between two features to build a compound feature. In Eq. (13), another mathematical operation binary log is applied to a single value of two features separately and after that we add them for making compound feature.

$$S = \sum_{k=1}^{n}(a_k + b_k) \quad (11)$$

$$M = \prod_{k=1}^{n}(a_k b_k) \quad (12)$$

$$Y_1 = \log_2 a_k, Y_2 = \log_2 b_k \quad (13)$$

We have considered raw features as the first level and the others as are inner level. Fig. 8 shows the level indexing process.

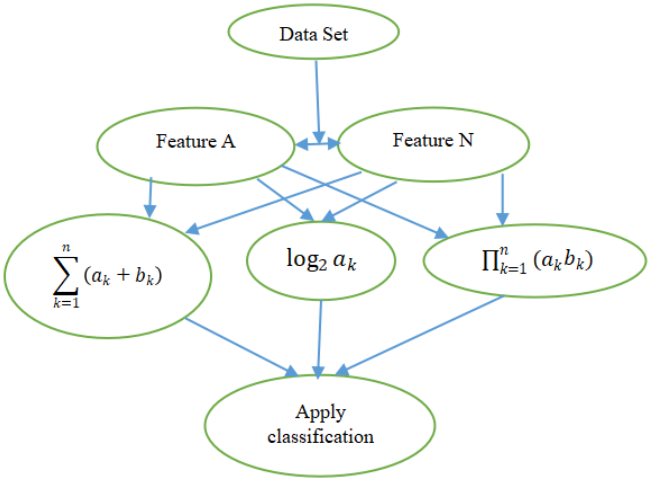

Fig. 8. Inner level processing system

VI. Results and Discussion

To evaluate the generalization of the model we have considered K fold cross-validation for low bias and a modest variance [33]. We have used five folds for training and testing the dataset and obtained the mean performance.

For the classification of the driver, we have introduced the prominent supervised algorithm named Naive Bayes (NB), Logistic Regression, *k*-NN, Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), and SVM. Table I shows the result of all classifiers through the confusion metric. Among them, GB shows 83.00 % (highest) and NB performs 65.00% (lowest) accuracy respectively. Mentionable, we have made four compound features from two features among 51 features and considered two drivers among 10 drivers respectively from the Ocslab dataset. The results of non-linear classifiers where GB and RF are given 85.00% and 81.00% accuracy accordingly. Results of confusion metrics are shown in Table II and a comparative analysis of original features vs compound features is visible in Table III. In the Compound feature, GB performs a height accuracy of 85%. Moreover, the ROC curve are shown in true positive in Fig. 10 for all classifiers.

Again we have considered the time series of the drivers' and found out the driving pattern style through the windowing system using the compound dataset. Fig. 9 is representing the accuracy of only two drivers (A, D) and it also represents the W_1, W_2, and W_3 accuracy of 75.00%, 78.00%, and 85.00% respectively. In this research, we have figured out linear and non-linear algorithms to calculate the performance. Moreover, according to pair plot analysis, we have decided that non-linear is better than linear analysis. As a result, the model is statistically significant because of the better performance of the non-linear algorithm.

TABLE I. Model Classification Report

Classifiers	***Performance parameters***				
		Accuracy	Precision	Recall	F1-Score
Logistic Regression		72.00	76.00	81.00	78.00
	Macro Average	-	71.00	70.00	70.00
	Weighted Average	-	72.00	72.00	72.00
Gradient Boosting		85.00	94.00	74.00	83.00
	Macro Average	-	81.00	83.00	81.00
	Weighted Average	-	84.00	81.00	81.00
Decision Tree		78.00	97.00	66.00	78.00
	Macro Average	-	80.00	81.00	78.00
	Weighted Average	-	84.00	78.00	78.00
Random Forest		81.00	86.00	80.00	83.00
	Macro Average	-	79.00	80.00	79.00
	Weighted Average	-	81.00	80.00	80.00
SVM		76.00	85.87	74.00	79.00
	Macro Average	-	75.00	76.00	75.00
	Weighted Average	-	77.00	76.00	76.00
k-NN		78.00	85.99	78.99	81.99
	Macro Average	-	76.00	77.00	77.00
	Weighted Average	-	78.00	77.00	78.00
Naïve Bayes		65.00	93.00	76.00	84.00
	Macro Average		57.00	64.00	57.00
	Weighted Average		85.00	74.00	78.00

TABLE II. Confusion Metrics

Classifiers	***Array metrics***	
Logistic Regression	1189	287
	379	539
Gradient Boosting	1097	379
	72	846
Decision Tree	1076	4000
	33	885
Random Forest	1183	293
	187	731
SVM	1093	383
	200	718
k-NN	1147	329
	210	708

TABLE III. Comparative Analysis of Original Feature and Compound Feature based on Model Accuracy

Classifiers	***Original feature***	***Compound Feature***
Logistic Regression	55	72
Gradient Boosting	80	85
Decision Tree	76	78
Random Forest	81	80
SVM	74	76
k-NN	76	78
Naïve Bayes	57	65

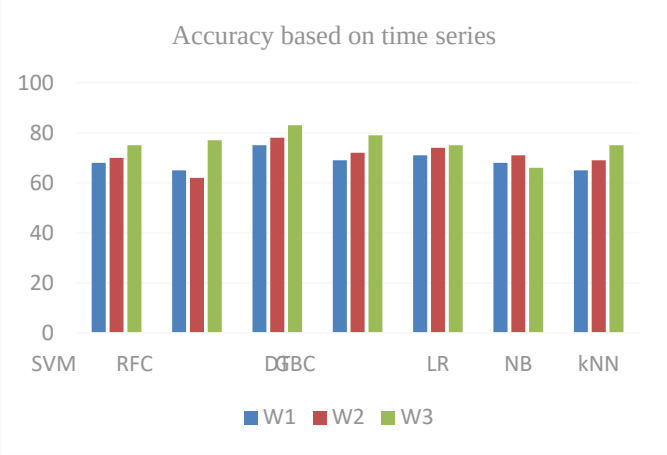

Fig. 9. Accuracy based on time series

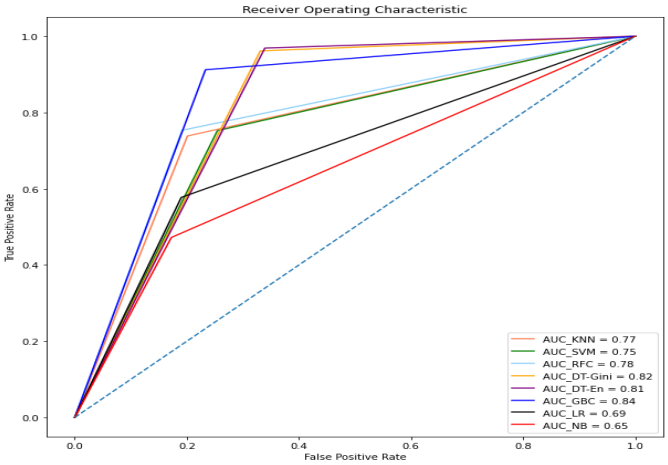

Fig. 10. ROC curve of the several classifiers

VII. Conclusion

In terms of driver identification, the role of machine learning as a cutting-edge technology is immense. Through this work, we found the best model that shows a momentous way to build our proposed anti-theft driving expert system. This research work has presented an in-depth comparison of the performance of seven (7) projecting classifiers in the context of driving behavior identification. As a result, the best and the worst classifiers we have found are GB and NB respectively. This type of result is very effective in developing an online-based anti-theft system. We have executed our experiment with only a few numbers of drivers among ten drivers.

Through online the owner of the vehicle will notify, so there is possibility man-in-the middle attack to another LAN. We will introduce network security to ensure authorized access and overcome the vulnerabilities of the proposed system in the next work.

Acknowledgment

I am grateful to KIA Motor Corporation for the Dataset provided to me.

References

[1] Enev M, Takakuwa A, Koscher K, Kohno T. Automobile Driver Fingerprinting. Proc. Priv. Enhancing Technol.. 2016 Jan; 2016(1):34-50.

[2] K. Uvarov and A. Ponomarev, "Driver Identification with OBD-II Public Data," 2021 28th Conference of Open Innovations Association (FRUCT), 2021, pp. 495-501, doi: 10.23919/FRUCT50888.2021.9347648.

[3] Kaplan S, Guvensan MA, Yavuz AG, Karalurt Y. Driver behavior analysis for safe driving: A survey. IEEE Transactions on Intelligent Transportation Systems. 2015 Aug 26; 16 (6):3017-32.

[4] Oak Ridge National Laboratory for the U.S. Department of Energy and the U.S. Environmental Protection Agency (U.S Department of Energy)

[5] https://www.businesswire.com/news/home/20190523005089/en/Worldwide-Connected-Vehicle-Shipments-Foreca St-to-Reach-76-Million-Units-by-2023-According-to-IDC (Accessed date 9/20/2022)

[6] General Motors. "Chevrolet Malibu Media Archives." (June 1, 2011) http://archives.media.gm.com/division/2003_prodinfo/03_chevrolet/03_malibu/index.html

[7] Jamie Page Deaton "5 Ways Modern Car Engines Differ from Older Car Engines" 16 May 2011. HowStuffWorks.com.<https://auto.howstuffworks.com/5-ways-modern-car-engines-differ-from-older-car-engines.htm> (accessed date 20 August 2022)

[8] Carfora MF, Martinelli F, Mercaldo F, Nardone V, Orlando A, Santone A, Vaglini G. A “pay-how-you-drive” car insurance approach through cluster analysis. Soft Computing. 2019 May; 23(9):2863-2875.

[9] Kwak BI, Woo J, Kim HK. Know your master: Driver profiling-based anti-theft method. In2016 14th Annual Conference on Privacy, Security and Trust (PST) 2016 Dec 12 (pp. 211-218). IEEE.

[10] “dailymail, ”http://www.dailymail.co.uk/news/article-2938793/Carhackers-driving-motors-Increasing-numbers-stolen-thieves-simplybypass-security-devices.html, 2015, (accessed: 2022-20-01).

[11] Penny Hoelscher “Information security-what is relay attack” january 31, 2019

[12] Pekaric I, Sauerwein C, Haselwanter S, Felderer M. A taxonomy of attack mechanisms in the automotive domain. Computer Standards & Interfaces. 2021 Apr 23:103539.

[13] BMW center, www.bmwusa.com/ConnectedDrive (accessed date 21-01-2022)

[14] Grant Maloy Smith ,https://dewesoft.com/author/grant-maloy-smith (accessed date 21-01-2022)

[15] Comparison of Event-Triggered and Time-Triggered Concepts with Regard to Distributed Control Systems A. Albert, Robert Bosch GmbH Embedded World, 2004, Nürnberg

[16] Ullah S, Kim DH. Lightweight driver behavior identification model with sparse learning on In-Vehicle CAN-BUS sensor data. Sensors. 2020 Jan;20 (18):5030.

[17] Girma A, Yan X, Homaifar A. Driver identification based on vehicle telematics data using lstm-recurrent neural network. In2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) 2019 Nov 4 (pp. 894-902). IEEE

[18] Wakita T, Ozawa K, Miyajima C, Igarashi K, Itou K, Takeda K, Itakura F. Driver identification using driving behavior signals. IEICE TRANSACTIONS on Information and Systems. 2006 Mar 1;89 (3):1188-94.

[19] Jun X, Zhao X, Rong J. A study of individual characteristics of driving behavior based on hidden markov model. Sensors & Transducers. 2014 Mar 1;167 (3):194.

[20] Castignani G, Derrmann T, Frank R, Engel T. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. IEEE Intelligent transportation systems magazine. 2015 Jan 19;7 (1):91-102.

[21] Kashevnik A, Lashkov I, Gurtov A. Methodology and mobile application for driver behavior analysis and accident prevention. IEEE transactions on intelligent transportation systems. 2019 Jun 4;21 (6):2427-36.

[22] Van Ly M, Martin S, Trivedi MM. Driver classification and driving style recognition using inertial sensors. In 2013 IEEE Intelligent Vehicles Symposium (IV) 2013 Jun 23 (pp. 1040-1045). IEEE.

[23] Enev M, Takakuwa A, Koscher K, Kohno T. Automobile Driver Fingerprinting. Proc. Priv. Enhancing Technol.. 2016 Jan;2016 (1):34-50.

[24] Choi S, Kim J, Kwak D, Angkititrakul P, Hansen JH. Analysis and classification of driver behavior using in-vehicle can-bus information. InBiennial workshop on DSP for in-vehicle and mobile systems 2007 Jun (pp. 17-19).

[25] Kedar-Dongarkar G, Das M. Driver classification for optimization of energy usage in a vehicle. Procedia Computer Science. 2012 Jan 1; 8:388-93.

[26] Xun Y, Liu J, Kato N, Fang Y, Zhang Y. Automobile driver fingerprinting: A new machine learning based authentication scheme. IEEE Transactions on Industrial Informatics. 2019 Oct 10;16(2):1417-26.

[27] Girma A, Yan X, Homaifar A. Driver identification based on vehicle telematics data using lstm-recurrent neural network. In2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) 2019 Nov 4 (pp. 894-902). IEEE

[28] Driving Dataset. Available online: https://ocslab.hksecurity.net/Datasets/driving-dataset (accessed on 2-11 2022)

[29] OBD-II PIDs, Available online: https://en.wikipedia.org/wiki/OBD-II_PIDs (accessed on 22-10-2022)

[30] Kwak, Sang Kyu, and Jong Hae Kim. "Statistical data preparation: management of missing values and outliers." Korean journal of anesthesiology 70.4 (2017): 407-411.

[31] Caruso, Pier Francesco, et al. "The effect of COVID-19 epidemic on vital signs in hospitalized patients: a pre-post heat-map study from a large teaching hospital." Journal of clinical monitoring and computing 36.3 (2022): 829-837.

[32] Chen, Tong, Yong-Xin Liu, and Luqi Huang. "ImageGP: An easy-to-use data visualization web server for scientific researchers." iMeta 1.1 (2022): e5.

[33] Jason Brownlee, https://machinelearningmastery.com/k-fold-cross-validation (accessed on 11-02-2022)

[34] Mo, Yujian, et al. "Review the state-of-the-art technologies of semantic segmentation based on deep learning." Neurocomputing 493 (2022): 626-646..

[35] M. T. Habib, A. Majumder, A. Jakaria, M. Akter, M. S. Uddin, & S. Ahmed, (2020). Machine vision based papaya disease recognition. Journal of King Saud University - Computer and Information Sciences, 32(3), 300– 309

[36] Rafatirad, Setareh, et al. "What Is Applied Machine Learning?." Machine Learning for Computer Scientists and Data Analysts. Springer, Cham, 2022. 3-33.

[37] Raju, VN Ganapathi, et al. "Study the influence of normalization/transformation process on the accuracy of supervised classification." 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2020.

[38] Gu Z. Complex heatmap visualization. iMeta. 2022 Sep;1(3):e43.

[39] Pitroda H. A Proposal of an Interactive Web Application Tool QuickViz: To Automate Exploratory Data Analysis. In2022 IEEE 7th International conference for Convergence in Technology (I2CT) 2022 Apr 7 (pp. 1-8). IEEE.