Fuel Consumption Prediction Model using Machine Learning

Mohamed A. HAMED, Mohammed H.Khafagy, Rasha M.Badry
Department of Information Systems, Faculty of Computers and Information
Fayoum University, Fayoum 63514, Egypt

***Abstract*—In the paper, we are enhancing the accuracy of the fuel consumption prediction model with Machine Learning to minimize Fuel Consumption. This will lead to an economic improvement for the business and satisfy the domain needs. We propose a machine learning model to predict vehicle fuel consumption. The proposed model is based on the Support Vector Machine algorithm. The Fuel Consumption estimation is given as a function of Mass Air Flow, Vehicle Speed, Revolutions Per Minute, and Throttle Position Sensor features. The proposed model is applied and tested on a vehicle's On-Board Diagnostics Dataset. The observations were conducted on 18 features. Results achieved a higher accuracy with an R-Squared metric value of 0.97 than other related work using the same Support Vector Machine regression algorithm. We concluded that the Support Vector Machine has a great effect when used for fuel consumption prediction purposes. Our model can compete with other Machine Learning algorithms for the same purpose which will help manufacturers find more choices for successful Fuel Consumption Prediction models.**

Keywords*—*Fuel consumption; machine learning; support vector machine; feature weight; feature selection; on-board diagnostic

ABBREVIATIONS

DT:	***Decision Tree***
FC:	***Fuel Consumption***
FS:	***Feature Selection***
GB:	***Gradient Boosting***
IoT:	***Internet of Things***
ML:	***Machine Learning***
MAF:	***Mass Air Flow***
MAE:	***Mean Absolute Error***
NN:	***Neural Networks***
OBD:	***On-Board Diagnostics***
RF:	***Random Forest***
RFE:	***Recursive Feature Elimination***
RMSE:	***Root Mean-Squared Error***
RBF:	***Radial Basis Function***
RPM:	***Revolution Per Minute***
SVM:	***Support Vector Machine***
ANN:	***Artificial Neural Network***
TPS:	***Throttle Position Sensor***
VS:	***Vehicle Speed***
ECU:	***Electronic Control Units.***

I. INTRODUCTION

In this study, we are trying to enhance fuel consumption (FC) prediction using machine learning algorithms. We used a Support Vector Machine algorithm to predict fuel consumption. We measure fuel consumption based on a legacy Dataset containing On-Board Diagnostics (OBD) data. The aim is to achieve a good value for the R-Squared metric using the SVM.

OBD is the protocol responsible for scanning and reading the ECU in the vehicle. OBD adapter can scan the ECU and send the FC data to a third-party device. OBD is considered a part of the Internet of Things technique. It can be connected to remote datasets to save its data for important and urgent analysis related to vehicles depending on Big Data, Deep Learning, and Machine Learning techniques. These analyses are helpful for instant diagnoses for vehicles and other types of machines which are using the same OBD protocol[1-4].

Fuel Consumption has an essential interest for individuals, businesses, and the globe. The price of fuel controls the economy of the world. Therefore, changes in the price of fuel affect the economical side for businesses.

Machine Learning is considered an application of Artificial Intelligence. Arthur Samuel said that Machine Learning: "is defined as the field of study that gives the computers the ability to learn without being explicitly programmed" [5].

One of the famous algorithms of Machine Learning is the Support Vector Machine (SVM) algorithm. SVM is an algorithm that tries to predict a specific value or a set of classes either in classification or regression form [6, 7]. It has been used in several studies related to the prediction of fuel consumption. These studies are considered to be related to our work similarly.

We used SVM to propose an ML model for fuel consumption prediction purposes. The other related research work had applied the SVM algorithm to predict FC based on a training dataset of a small size. Its results were not enough good. Its model had returned an R-Squared value equal 0.004624. It depended on the RPM_TPS-based equation only, which will be discussed later. However, in our research, we used both the RPM_TPS-based equation besides the VS_MAF-based equation. There is no other literature that discussed the same problem with the SVM algorithm depending on both RPM_TPS-based and VS_MAF-based equations. The RPM_TPS-based equation depends on RPM and TPS parameters. The VS_MAF-based equation depends on VS and

MAF parameters. These two equations are considered the most important equations that can be used to measure the fuel consumption rate when a complete FC Dataset exists. Our FC Dataset is considered a high-dimensional size dataset.

It's important to note that our proposed model and its internal experiments couldn't be observed without an FC Dataset containing the parameters which are existing in the FC equations used.

Before using SVM for the prediction of fuel consumption, Feature Weighting should be described. Feature Weighting is the ranking process of the importance of the features, as it depends on a voting approach for ranking the importance of the features in datasets [8].

Feature Weighting is followed by the Feature Selection step. Feature Selection is applied to the highly ranked features after the Feature Weighting step. Then, these highly ranked features are filtered and applied to the classifier [8].

Feature Selection can be applied to datasets using different algorithms. Random Forest and Decision Tree are the most famous algorithms used to rank the importance of the features and select the highly ranked features.

In the last decade, scholars talked about the importance of predicting the consumed fuel percentage depending on some of the sophisticated algorithms from both Data Mining (DM) and Machine Learning (ML). However, in an earlier time, scholars had discussed the prediction of fuel consumption with different algorithms, including Neural Networks (NN), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM) [9, 10].

The prediction of fuel consumption value will become more precise when predicted with sophisticated ML techniques. The discussion of fuel consumption has been a trending topic when discussed from the view of ML in the last five years.

Many research papers have been developed to discuss the most followed methods for monitoring fuel consumption in vehicles. Fuel consumption scholars have focused on different methods that should be followed to eliminate fuel consumption.

In [11], the authors had used sophisticated techniques depending on ML models to detect and measure levels of fuel consumption using Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) models. They used 27 vehicles in their experiments. They discussed their multiple tries for achieving better accuracy on different types of vehicles of the same age, different segments, engine displacement, and type of transmission. Finally, they achieved accuracy with 83%.

In [12], the authors had discussed the problem of predicting fuel in fleets of vehicles depending on machine learning techniques. They had used Random Forest, Gradient Boosting, and Neural Networks as machine learning models. Random Forest Algorithm had achieved the best result between the other used algorithms. However, they depended on the Nash-Sutcliffe coefficient for measuring the predictive power for the efficiency of each model. Also, they used Bias, Mean Absolute Error (MAE), and Root Mean-Squared Error (RMSE) as error statistics to evaluate their model's accuracy.

In [13], the authors had used a machine-learning algorithm to predict fuel consumption depending on a set of variables in a large-scale Dataset gathered by 153 drivers during a month depending on GPS and CAN (Controller Area Network) bus data, including speed of the vehicle and moved distance. They used regression methods for the machine learning methods: SVM, ANN, Linear Regression (LR), and Link Fuel Summation SVM model (LSSVM). Their study revealed that SVM had the best R-Squared value with 0.92 while ANN, LR, and LSSVM had R-Squared values of 0.86, 0.74, and 0.79. The training phase had affected the superiority of SVM over other models. However, SVM had generated the best fit results/accuracy. Also, it wasn't affected by cost functions as it provided a linear penalty to huge error rates where the ANN model minimizes the sum of squared errors.

In [14], the authors had used Boruta Algorithm (BA) and Neural Networks (NNs) algorithm to measure fuel consumption regarding a huge fleet of trucks on different road pavements. BA had shown a good result in comparison with previous studies, which used the same data. While the developed NN algorithm had achieved (R2) value of 0.88 for test data. NN appeared to be a suitable candidate for analyzing large datasets effectively and predicting the impact of roughness and macrotexture of roads on truck fuel consumption.

In [15], the authors had addressed the identification of driving style issues. They used the K-means clustering algorithm to differentiate between different types of driving styles. Driving styles are divided into three categories: normal, soft, and aggressive category. Also, they used random forest, K-nearest neighbor, support vector machine, and neural network models. Random forest overall accuracy was 95.39% while trucks are in their heavy load, and 90.74% on no-load status. The aggressive driving style achieved the largest fuel consumption and reached 10 % higher than the average driving style.

In [16], the authors had used Autonomie, which is a simulation tool, to simulate the process of fuel and vehicle power consumption. They proposed a Large-scale learning and prediction process (LSLPP) with machine learning models. LSLPP tests were successful as they could accelerate analysis processes and prediction of vehicle's fuel consumption.

In [17], the authors had used the Support Vector Machine (SVM) model as one of the ML prediction techniques with OBD-II to monitor and predict fuel consumption levels. The proposed model uses both TPS and RPM variables to measure the consumed level of fuel. Finally, their RMSE value was 2.43.

In [18], the authors had used SVM, RF, and ANN algorithms for fuel consumption prediction purposes. SVM and ANN algorithms achieved the best results. However, RF outperformed both of them. The coefficient of determination (R2) for SVM, RF, and ANN are 0.83, 0.87, and 0.85, respectively.

II. PROPOSED MODEL

The proposed model aims to predict fuel consumption using SVM. The proposed model consists of four phases: Data Preprocessing, Feature Weighting, Feature Selection, and SVM Prediction Model, as shown in Fig. 1. The proposed prediction model has been applied to FC Dataset with 8262 records. The Dataset includes 18 fields, as shown in Table I. FC Dataset was gathered by 19 drivers using an OBD scanner in vehicles, which was used for a previous dissertation for profiling automotive data in 2018 [19]. The Dataset gathered by 19 drivers had been collected depending on a vehicle model of the well-known Brazilian vehicle, A 2015 Chevrolet S10, which has a 2.5-liter flex-fuel engine by 206 hp. This Dataset is gathered in an urban road in the city of Natal (Brazil). It was gathered at a distance of 18.8 kilometers for 34 minutes for each driver [20, 21].

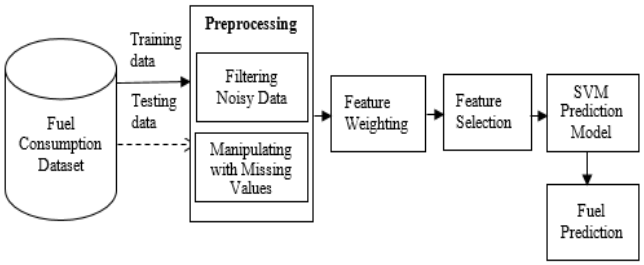

Fig. 1. Proposed Model Diagram.

TABLE I. FC DATASET FIELDS

No.	Field Name	Field Description
1	TIME	Vehicle's work period.
2	LATITUDE	Latitude of the vehicle while reading time.
3	LONGITUDE	Longitude of the vehicle while reading time.
4	ALTITUDE	The altitude of the vehicle while reading time.
5	BAROMETRIC_ PRESSURE	Measuring the atmospheric pressure.
6	ENGINE_COOLANT_ TEMP	Measuring the engine's temperature.
7	FUEL_LEVEL	Level of fuel in tank at the reading time.
8	ENGINE_LOAD	Measure sucked air and fuel into the engine.
9	AMBIENT_AIR_TEM P	Refer to the outside air temperature.
10	ENGINE_RPM	Refer to the frequency of rotation around a fixed axis.
11	INTAKE_MANIFOLD _PRESSURE	Refer to the negative air pressure inside the intake pipe.
12	MAF	Measure the amount of air entering the engine.
13	AIR_INTAKE_TEMP	Refer to air temperature in the engine.
14	SPEED	Speed value of the vehicle at request time.
15	Short Term Fuel Trim Bank 1	Refer to ECU signaling response according to the changes of the oxygen levels.
16	THROTTLE_POS	Identify the value of air-delivered quantity to vehicle's engine accurately.
17	TIMING_ADVANCE	Refer to the required time for the air-fuel mixture to be burned.
18	EQUIV_RATIO	Refer to the commanded air/fuel ratio of the engine.

A. Pre-processing

Data pre-processing is the first step in the proposed prediction model. Converting the data into a more desired and eligible form is essential to ensure that the Dataset is accurate and ready for further processing [22, 23]. In our proposed model, we are performing filtering noisy data and manipulating with missing values steps.

1) Filtering noisy data: This step is a very important step in which the noisy records in FC Dataset are removed. For example, some cells are filled with symbols and characters like the Speed field, which contains (50 km/h). Such characters and symbols affect the implementation and results of the prediction model.

2) Manipulating with missing values: Most of the fields in our FC Dataset had filled with data. However, our FC Dataset was high-dimensional. Hence, it was difficult to discover the missing values by hand. So, we had to automate this process using specific techniques to avoid exceptions happening while training the SVM algorithm as the missing values may cause a big issue for processing the prediction model. For example, the FC Dataset contains fields with Nan, Null, or Zero values, in which the R2 value of the regression model is affected negatively and returned exceptions in the runtime. There are several methods to handle missing values. One of these methods is the mean imputation. The imputation method estimates the missing values by replacing them with the mean for that variable [24].

B. Feature Weighting

Feature weighting is an essential step in identifying the most feature or a set of features affecting other specific features. In the proposed model, feature weighting is used to set weights for FC Dataset features to identify which feature mostly affects the fuel consumption level. We used two models for weighting features in our Dataset. These models are Random Forest (RF) and Decision Tree Algorithm.

The Random Forest Algorithm is considered a good and reliable algorithm for features ranking for small and larger datasets. This is because it can distinguish the relevant and the irrelevant attributes in the Dataset. It can handle both classification and regression problems by constructing multiple decision trees concurrently and returning equivalent forecasting for the average result of the processed decision trees. RF can handle high-dimensional datasets as it can process too many inputs and return results with high performance [25, 26].

Decision Tree Algorithm is also an important model used to identify the importance of the attributes in the Dataset in which feature selection can be used in higher and lower-dimensional classification tasks. It describes the relations between data that can be simulated by leaves in the trees. Each node has other leaf nodes under it. Each leaf node holds a specific value that represents a meaningful form for the algorithm. A tree constitutes a leaf and a node. In classification, nodes represent a group to be classified and each node subset represents a value that can be taken by the node [27, 28].

Feature weighting is applied to the whole selected features of the FC Dataset to determine the most important features that affect fuel consumption. Clarification of feature weighting phase and used methodologies and algorithms will be discussed in another following section.

Feature weighting is applied to features in equations that are used for calculating fuel consumption. Fuel consumption can be calculated via two methods, the first is based on VS and MAF features, and the second is based on RPM and TPS features.

1) VS_MAF-based: VS_MAF is the first method used for calculating fuel consumption, according to (1).

$$f = VS / MAF \quad (1)$$

Where f is fuel consumption value, VS is the vehicle speed parameter, which is measured in km/hour, and MAF refers to the value of Mass Air Flow in the engine, which is measured in g/s (gram per second).

Depending on (1), fuel consumption can be measured using two metrics, the first metric is Mile Per Gallon (MPG), and the second metric is Liters per 100 Km. For example, the following equation retrieves the fuel consumption values in MPG and L/100KM.

To retrieve the fuel consumption value in US MPG, based on (2), the value of Speed is divided by MAF then multiplied by $\alpha = 7.718$, which is a constant.

$$f = VS * \alpha / MAF * \beta \quad (2)$$

Further, to retrieve fuel consumption value in liters per 100 km, we multiply the fuel consumption value in US MPG by the value of the constant β [17].

2) RPM_TPS-based: RPM_TPS is the second method used for calculating fuel consumption, according to (3).

$$Fuel_{(rpm, tps)} = p00x^2 + p10x + p01xy \quad (3)$$

Where X refers to Revolutions Per Minute (RPM), Y refers to Throttle Position Sensor (TPS), and the coefficients p00, p10, p01 values are 2.685, -0.1246, and 1.243, respectively.

Our feature selection experiments had been applied using the VS_MAF-based Equation and the RPM_TPS-based Equation. Generated results from our Random Forest Algorithm indicated that RPM, SPEED, and MAF have the highest effects on fuel consumption levels. Results and analysis for applying RF to FC Dataset will be provided in more detail in specific sections for the experiments, discussion, and results.

C. Feature Selection

After feature weighting, feature selection is applied to determine the most weighted features that affect the fuel consumption value after feature weighting. First, the generated weights of the FC features by the weighting models RF and DT are ranked, then a selection of the most important features is done.

D. The SVM Prediction Model

The proposed model is based on Support Vector Machine (SVM). The SVM model is a machine learning algorithm that reads input data and represents points on a 2d space or 3d space to be drawn on X-axis and Y-axis in 2d view or X-axis, Y-axis, and Z-axis in 3d view. Then, it draws a boundary line that splits the groups and classifies the data to refer to which class the point is grouped or classified. SVM has a maximum margin line that usually divides the class of points equally called "hyperplane". The hyperplane looks for the maximum distance between each point and its nearest group or class [18]. The hyperplane is divided into two different types. The first one is the optimal hyperplane, which is the linear function with the maximum margin between vectors or multiple vectors in two groups, and the second one is called the soft margin hyperplane, which happens when two classes of the data are not linearly separable [6].

E. Experiments

In the experiment section, the details of the performed experiments are illustrated. Several experiments had been done using a historical FC Dataset. However, the proposed work for predicting fuel consumption using SVM with a regression model is considered the first experiment with this algorithm to be conducted specifically on this Dataset. The experiments are conducted using two equations. The first is based on MAF and VS features, and the second is based on RPM and TPS features.

The results of the experiments have been evaluated using the coefficient of determination metric R-Squared/R^2, a statistical metric that represents the variance between dependent and independent variables and evaluates the model's ability for prediction purposes [29].

After applying feature weighting and selection, we update both the VS_MAF-based equation and the RPM_TPS-based equation. These updates improve the Squared Correlation Coefficient metric R-Squared/R^2 value of the proposed model compared with other studies.

1) Applying feature weighting: Feature Weight/importance is identified via different algorithms used to select the most important features in high-dimensional datasets. RF and DT are two important algorithms used to measure features and find the correlations in FC Dataset.

Random Forest (RF) is a machine learning algorithm commonly used to evaluate the model's ability for prediction purposes.

Recursive Feature Elimination (RFE) was first used and proposed to enable the SVM model to evaluate the features/attributes importance and identify field ranking in datasets. The same methodology has been added to the RF algorithm to find the correlated features/fields in datasets with high dimensionality [30].

RF and DT algorithms are reliable enough to be considered for measuring the importance of the features in our FC Dataset. Using RF and DT for features weighting purposes during our observation leads to a better focus of the prediction purpose, after removing the unnecessary features from the experiment.

We had imported both RF and DT algorithms in Spider engine to run them using Python v.9 programming language.

Python has become an essential programming language for ML research. We used Python to print the weight results for FC Dataset features. We could draw figures using Matplotlib, which is a drawing and visualization library using Python, to differentiate the features with high and low importance values [31].

We had applied the RF-RFE algorithm on most of the features of the FC Dataset, which are 18 features. We had removed 15 features from the original FC features, which were 33 features. For example, Term Fuel Trim Bank 1, FUEL_ECONOMY, Long Term Fuel Trim Bank 2, FUEL_TYPE, FUEL_PRESSURE, Short Term Fuel Trim Bank 2, and TROUBLE_CODES had been removed because they were empty field. DTC_NUMBERS had been removed because it contained String values like 4101000761001:00410100076100, which is not meaningful. ENGINE_RUNTIME had been removed because it contained time values like 12:03:20 AM. VEHICLE_ID had been removed because it contained String values like s11. Finally, we had applied the feature selection experiment on both the VS_MAF-based equation and the RPM_TPS-based equation. We had measured the importance/weight score of all measured features in our FC Dataset. The results of feature weighting algorithms which were returned with high importance values looked to exist in the FC equations that we are already depending on during our proposed model. That means that using RF and DT for feature weighting has returned reasonable features for measuring FC Dataset weights.

a) Feature Selection experiment applied on 18 features using (RF): We applied the Random Forest Algorithm for identifying the importance/weight score for the features existing in our Dataset. Table II shows the importance of our Dataset features. Fig. 2 and Fig. 3 show a representation of the feature's importance in our Dataset. It was found that the most important features that affect the fuel consumption level after applying the feature selection algorithm according to the VS_MAF-based equation are MAF and SPEED. However, when applying the RPM_TPS-based equation, it was found that RPM is the most important feature.

Fig. 2 indicates that MAF and SPEED parameters are the most important parameters in the Dataset that affect fuel consumption according to VS_MAF-based equations. While Fig. 3 indicates that ENGINE_RPM is the most important feature that affects fuel consumption between the whole features in the dataset according to the RPM_TPS-based equation.

b) Weighted VS_MAF-based equation: According to the VS_MAF-based equation, fuel consumption calculation is based on MAF and VS features. Depending on RF and DT algorithms, Table III, Fig. 4, and Fig. 5 represent the feature importance results for both MAF and VS features.

In Table III, and Fig. 4 the results show and indicate that both MAF and SPEED features affect fuel consumption features with a feature weight of 0.50876 for MAF and 0.49124 for SPEED using the RF algorithm. However, in Table III and Fig. 5 both the MAF and SPEED features affect the fuel consumption feature with a feature weight of 0.50665 for MAF and 0.49335 for SPEED using the DT algorithm. This indicates that the importance value of the MAF and SPEED features doesn't hugely change when applied to the RF or DT algorithms.

Also, the previous importance values for both of the features refer to the more significant impact of the MAF feature over the SPEED feature when compared to each other according to their effect on the fuel consumption value.

After calculating feature weight for MAF and VS, the VS_MAF-based equation can be updated by adding the weight values for the equation.

So, we can multiply each feature in the equation by its importance according to Table III to become:

$$f = VS * vs_i / MAF * maf_i \quad (4)$$

Where vs_i = 0.49124 and maf_i =0.50876 by RF algorithm.

TABLE II. WEIGHTS OF FEATURES USING RANDOM FOREST ALGORITHM

No.	Field Name	VS_MAF-Equation	RPM_TPS-Equation
1	TIME	0.00037	0.00000
2	LATITUDE	0.00046	0.00001
3	LONGITUDE	0.00046	0.00000
4	ALTIDUTE	0.00038	0.00001
5	BAROMETRIC_ PRESSURE	0.00018	0.00000
6	ENGINE_COOLANT_TEMP	0.00024	0.00000
7	FUEL_LEVEL	0.00028	0.00000
8	ENGINE_LOAD	0.00506	0.00001
9	AMBIENT_AIR_TEMP	0.00024	0.00000
10	ENGINE_RPM	0.00050	0.99992
11	INTAKE_MANIFOLD_PRESSURE	0.00035	0.00000
12	MAF	0.56186	0.00000
13	AIR_INTAKE_TEMP	0.00032	0.00001
14	SPEED	0.42643	0.00000
15	Short Term Fuel Trim Bank 1	0.00042	0.00000
16	THROTTLE_POS	0.00197	0.00000
17	TIMING_ADVANCE	0.00048	0.00000
18	EQUIV_RATIO	0.00000	0.00000

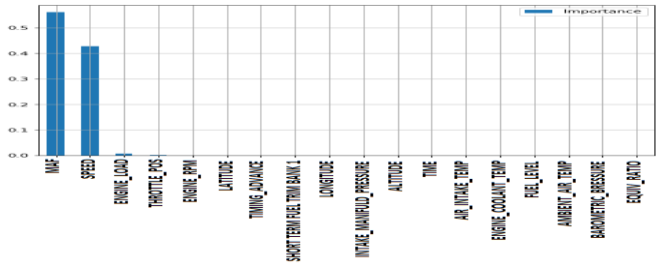

Fig. 2. Features Importance using VS_MAF-based Equation Results with Random Forest Algorithm.

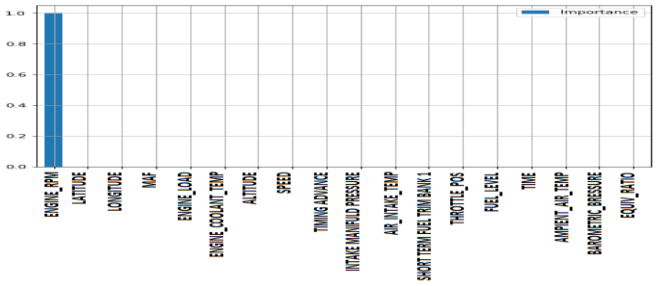

Fig. 3. Features Importance using RPM_TPS-based Equation Results with Random Forest Algorithm.

TABLE III. MAF AND VS FEATURE IMPORTANCE ACCORDING TO RF AND DT ALGORITHMS

VS_MAF-based Equation		
Algorithm	**MAF Weight**	**VS Weight**
Random Forest	0.50876	0.49124
Decision Tree	0.50665	0.49335

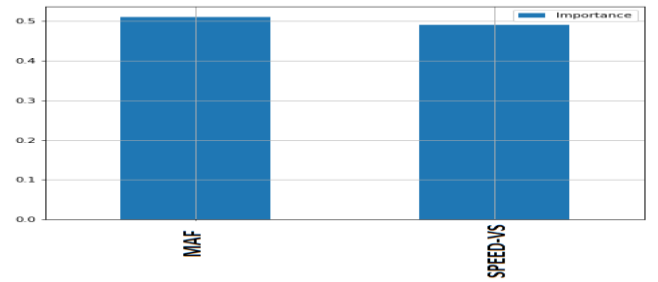

Fig. 4. MAF and VS Feature Importance using Random Forest Algorithm.

c) Weighted RPM_TPS-based equation: According to the RPM_TPS-based equation, fuel consumption calculation is based on RPM and TPS features. Therefore, depending on the RF and DT algorithms, Table IV and Fig. 6, and Fig. 7 include the feature importance results for both RPM and TPS features.

Table IV and Fig. 6 indicate that both RPM and TPS features affect the fuel consumption feature with a feature weight of 0.999952 for RPM and 0.000048 for TPS using the RF algorithm. However, in Table IV and Fig. 7, both the RPM and TPS features affect the fuel consumption with a feature weight of 0.999969 for RPM and 0.000031 for TPS using the DT algorithm. This indicates that the importance value of the RPM and TPS features doesn't change when applied to the RF or DT algorithms. Also, the previous importance values for both of the features refer to the massive importance of the RPM when compared with TPS importance.

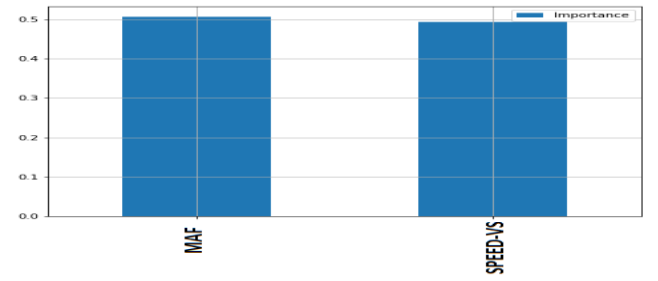

Fig. 5. MAF and VS Feature Importance using Decision Tree Algorithm.

TABLE IV. RPM AND TPS FEATURE IMPORTANCE ACCORDING TO RF AND DT ALGORITHMS

RPM_TPS-based Equation		
Algorithm	**RPM Weight**	**TPS Weight**
Random Forest	0.999952	0.000048
Decision Tree	0.999969	0.000031

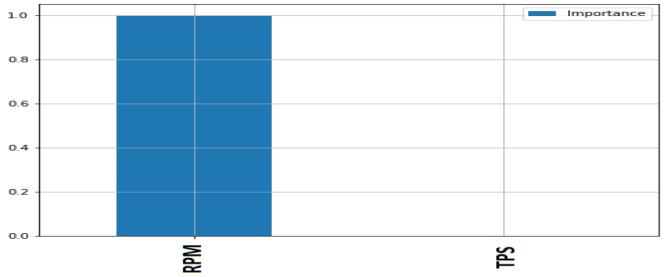

Fig. 6. RPM and TPS Features Importance using Random Forest Algorithm.

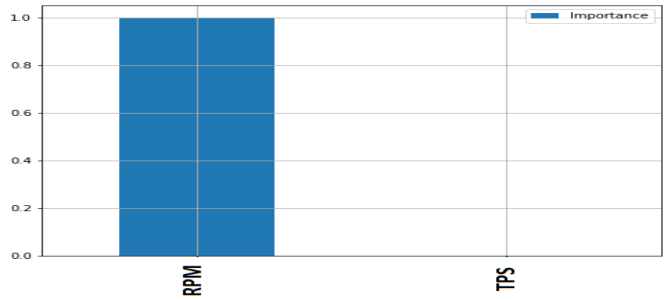

Fig. 7. RPM and TPS Features Importance using Decision Tree Algorithm.

The same as the VS_MAF-based equation, the RPM_TPS-based equation calculate fuel consumption rate using the following equation:

$$Fuel_{(rpm,\ tps)} = p00x^2 + p10x + p01xy \quad (5)$$

We can update the last equation via multiplying RPM and TPS by their importance values according to the generated results by the RF algorithm in Table IV to become:

$$Fuel_{(rpm,\ tps)} = p00x^2 * rpm_i + p10x * rpm_i + p01xy * rpm_i * tps_i \quad (6)$$

Where $rpm_i = 0.999952$ and $tps_i = 0.000048$ by RF algorithm.

2) Applying SVM on fuel consumption equations: The SVM model is applied using the original and the new-weighted fuel consumption equations, which calculates fuel consumption values. We had noticed the difference in the squared correlation coefficient $\mathbf{R^2}$ metric value for each conducted experiment.

Table V shows a sample of the data using the VS_MAF-based experiment and the RPM_TPS-based experiment. Fig. 8 compares the actual and predicted values of fuel using the VS_MAF-based equation when implemented using the SVM model, while Fig. 9 compares the actual and predicted values of fuel using the RPM_TPS-based equation when implemented using the SVM model.

In Fig. 8, according to the VS_MAF-based experiment, it looks that some of the actual fuel consumption data are quite

similar to the predicted values, which are likely similar to the result of the R-Squared/R^2 value of the model that reached 0.97, which indicates that the SVM model has achieved a high accuracy depending on the VS_MAF-based equation. Also, in Fig. 9, according to the RPM_TPS-based experiment, it looks that some of the actual data are quite similar to the predicted fuel consumption values, which are likely similar to the result of the R-Squared/R^2 value of the model that reached 0.96, which indicates that the SVM model has achieved a high accuracy too using the result of applying the RPM_TPS-based equation.

TABLE V. VS_MAF SAMPLE DATA

Dataset – Data Frame				
Index	**MAF (g/s)**	**SPEED (VS) (km/h)**	**RPM (rev/min)**	**TPS (%)**
0	24.77	48	2124	34.9
1	30.96	60	2617	36.1
2	18.58	64	3005	32.2
3	18.38	65	3156	32.5
4	19.77	67	1798	33.3
5	9.99	65	1818	28.6

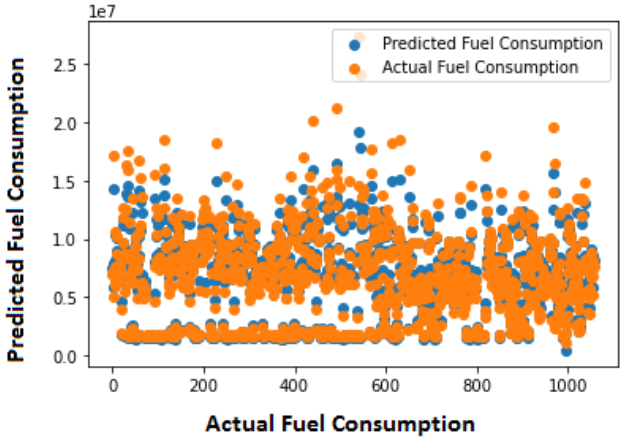

Fig. 8. Comparing the Actual to the Predicted Values of Fuel using VS_MAF-based Equation and SVM Algorithm.

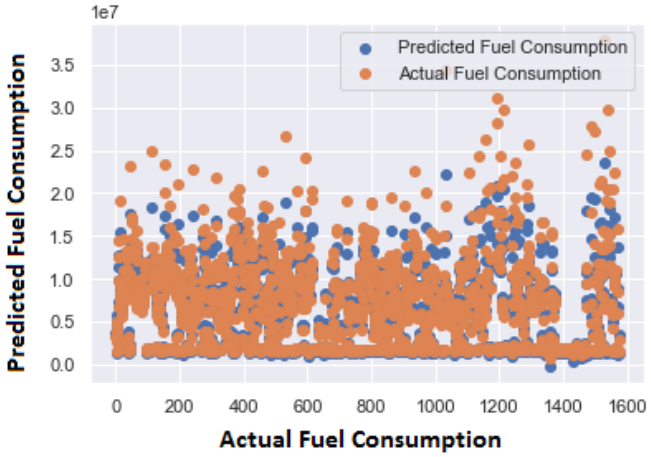

Fig. 9. Comparing the Actual to the Predicted Values of Fuel using RPM_TPS-based Equation and SVM Algorithm.

III. DISCUSSION

The performance of the proposed model is evaluated using the R-Squared/R^2 metric. The model is applied using two equations. The new VS_MAF-based equation depended on two variables are SPEED and MAF, while the new RPM_TPS-based equation depended on RPM and TPS variables. The squared correlation coefficient R-Squared/R^2 metric using the original VS_MAF-based equation and the original RPM_TPS-based equation is 0.96, according to Table VI.

We can notice that the squared correlation coefficient R-Squared/R^2 metric value has changed after applying the new-weighted equations for the VS_MAF-based equation. For example, the squared correlation coefficient R-Squared/R^2 metric using the new VS_MAF-based equation became 0.97 after applying the new-weighted equation for the original VS_MAF-based equation, according to Table VI. This is because we had depended on a Radial Basis Function (RBF) as a kernel function of the SVM model while training and testing phases of the VS_MAF-based equation. However, we had depended on a Linear function of the SVM model while the training and testing phases of the RPM_TPS-based equation, which became 0.96 after applying the new-weighted RPM_TPS-based equation.

We had achieved a superior result better than other candidates [17] who achieved lower results: R-Squared/R^2 = 0.004624 than our experiment using the same SVM predictor model depending on the RPM_TPS-based equation. Therefore, our goodness of fit using the R-Squared/R^2 metric equals 0.96 when applied their original RPM_TPS-based equation and our new-weighted RPM_TPS-based equation, according to Table VI.

Finally, the new weighted-VS_MAF-based equation had affected the R-Squared/R^2 metric value to be 0.97, while the new weighted-RPM_TPS-based equation had affected the R-Squared/R^2 metric value to be 0.96, according to Table VII.

IV. RESULTS

The value of the R-Squared/R^2 metric is 0.96 for the original VS_MAF-based equation and the original RPM_TPS-based equation, while the value of the R-Squared/R^2 metric is 0.97 while applying the New-weighted VS_MAF-based equation and 0.96 for applying the New-weighted RPM_TPS-based equation.

Table VI refers to the R-Squared/R^2 value of our prediction model, SVM, using the original and the new-weighted equations.

TABLE VI. R2 METRIC VALUE OF SVM MODEL WHEN APPLYING THE ORIGINAL AND THE NEW-WEIGHTED EQUATIONS FOR PREDICTION OF FUEL CONSUMPTION

FC equation	**R-Squared/R^2 metric value**		
	Proposed model results		**Related work results**
	VS_MAF	**RPM_TPS**	**RPM_TPS**
Original	0.96	0.96	0.004624
New-weighted	0.97	0.96	----------

TABLE VII. R2 METRIC VALUES AFTER APPLYING THE VS_MAF AND RPM_TPS EQUATIONS USING THE SVM MODEL

FC Equation	R-Squared (R^2)
Proposed weighted VS_MAF- equation	0.97
Proposed Weighted RPM_TPS- equation	0.96

Finally, Table VII shows the final result of the squared correlation coefficient R-Squared/R^2 metric value using the new weighted VS_MAF-based equation and new weighted RPM_TPS-based equation, respectively.

V. CONCLUSION

This study had applied four different equations in four experiments on a historical OBD dataset for predicting fuel consumption using the SVM regression model as an ML technique. These equations were the original VS_MAF-based equation and the new weighted-based equation which depended on Speed and MAF variables. Also, these equations included the original RPM_TPS-based equation and the new weighted-based equation which depended on RPM and TPS variables.

The squared correlation coefficient R-Squared/R2 metric value of the SVM model became 0.96 for both of the original equations, and (0.97, 0.96) for the new equations, VS_MAF-based equation, and RPM_TPS-based equation, respectively.

This study had achieved better results than other candidates who applied the RPM_TPS-based equation using the SVM model [17] as their R-Squared/R^2 equals: 0.004624 while implementing their RPM_TPS-based equation.

Future research may be a try to investigate more correlated parameters in FC Dataset to create more mathematical equations for measuring FC. The more correlated parameters the more equations that calculate FC, consequently, the more experiments and observations that lead to better enhancements and more accurate FC prediction models with ML. MAF, RPM, SPEED, ENGINE_LOAD, and ENGINE_RPM may have a greater correlation that can be used to create a new mathematical equation to measure FC in our FC Dataset or we can use a larger dataset for implementing FC prediction observations. Also, the great enhancement will be to convert the proposed model to a running system integrated with Internet of Things components and devices in the vehicle and predict fuel consumption instantly with SVM in the runtime applying our proposed model.

REFERENCES

[1] A. A. M. S. Medashe Michael Oluwaseyi "Specifications and Analysis of Digitized Diagnostics of Automobiles: A Case Study of on Board Diagnostic (OBD II)," vol. 9, 1 ed: IJERT, 2020.

[2] T. A. Nikolaos Peppes, Evgenia Adamopoulou and Konstantinos Demestichas, "Driving Behaviour Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data," vol. 4704, Machine Learning Applied to Sensor Data Analysis ed: MDPI: sensors, 2021.

[3] A. K. S. Siddhanta Kumar Singh, and Anand Sharma, "OBD - II based Intelligent Vehicular Diagnostic System using IoT," ed. ISIC'21: International Semantic Intelligence Conference: Mody University of Science and Technology, Lakshmangarh, Sikar, Rajasthan, India, 2021.

[4] U. M. P. Pirapuraj, "Intelligent Vehicle Diagnostic System for Service Center using OBD-II and IoT," ed. Conference: International Conference of Science and Technology - 2021: Faculty of Technology, South Eastern University of Sri Lanka, Oct 2021.

[5] B. Mahesh, "Machine Learning Algorithms - A Review," vol. 9, 1 ed: International Journal of Science and Research (IJSR), 2020, p. 7.

[6] C. Cortes, & Vapnik, V., "Support-vector networks. ," vol. 20, ed. Machine learning, 1995, pp. 273–297.

[7] M. P. Theodoros Evgeniou, "WORKSHOP ON SUPPORT VECTOR MA CHINES: THEORY AND APPLICATIONS," vol. 2049, ed. Conference: Advanced Course on Artificial Intelligence (ACAI 1999) :Machine Learning and Its Applications , Lecture Notes in Computer Science, 2001, pp. 249-257.

[8] M. D. d. L. N. L. da. Costa, R. Barbosa, "Evaluation of feature selection methods based on artificial neural network weights," vol. 168, ed. Expert Systems with Applications (2020), 2020.

[9] A. Schoen, Byerly, A., Hendrix, B., Bagwe, R. and d. S. M., E. C., & Miled, Z. B., "A machine learning model for average fuel consumption in heavy vehicles.," vol. 68, 7 ed. IEEE Transactions on Vehicular Technology: IEEE, 2019, pp. 6343-6351.

[10] Y. Yao, Zhao, X., Liu, C., Rong, J., Zhang, Y., and Z. Dong, & Su, Y., "Vehicle fuel consumption prediction method based on driving behavior data collected from smartphones," vol. 2020, ed. Journal of Advanced Transportation, 2020.

[11] E. Moradi, & Miranda-Moreno, L., "Vehicular fuel consumption estimation using real-world measures through cascaded machine learning modeling. ," vol. 88, ed. Transportation Research Part D: Transport and Environment, 2020.

[12] S. Wickramanayake, & Bandara, H. D., "Fuel consumption prediction of fleet vehicles using machine learning: A comparative study. ," ed. In 2016 Moratuwa Engineering Research Conference (MERCon) IEEE., 2016, pp. 90-95.

[13] W. Zeng, Miwa, T., & Morikawa, T., "Exploring trip fuel consumption by machine learning from GPS and CAN bus data. ," vol. 11, ed. Journal of the Eastern Asia Society for Transportation Studies, 2015, pp. 906-921.

[14] F. Perrotta, Parry, T., Neves, L. C., & and M. Mesgarpour, "A machine learning approach for the estimation of fuel consumption related to road pavement rolling resistance for large fleets of trucks.," ed. The Sixth International Symposium on Life-Cycle Civil Engineering (IALCCE 2018), 2018.

[15] Q. Wang, Zhang, R., Wang, Y., & Lv, S., "Machine learning-based driving style identification of truck drivers in open-pit mines. ," vol. 9, 1 ed: Electronics, 2020, p. 19.

[16] J. Yao, & Moawad, A., "Vehicle energy consumption estimation using large scale simulations and machine learning methods.," vol. 101, ed. Transportation Research Part C: Emerging Technologies, 2019, pp. 276-296.

[17] T. Abukhalil, AlMahafzah, H., Alksasbeh, M., and B. A. & Alqaralleh, "Fuel consumption using OBD-II and support vector machine model.," vol. 2020, ed. Journal of Robotics, 2020, pp. 1-9.

[18] F. Perrotta, Parry, T., & Neves, L. C., "Application of machine learning for fuel consumption modelling of trucks.," ed. In 2017 IEEE International Conference on Big Data (Big Data) IEEE, 2017, pp. 3810-3815.

[19] B. a. C. A. D. Silveira, "Use of machine learning techniques to identify car usage profiles based on automotive data," ed. https://repositorio.ufrn.br/jspui/handle/123456789/26017: Metropole Digital Institute, Fedral University of Rio Grande do Norte, Natal, 2018.

[20] J. C. Cephas A, Anne M, Ivanovitch S., "A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification.," ed. In Anais do XV Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)At: Brazil, 2018, pp. 174-185.

[21] C. A. d. S. Barreto, "OBD-II Datasets," no. v0.3, 2018. [Online]. Available: https://www.kaggle.com/cephasax/obdii-ds3.

[22] U. a.-l. a. c. r.-.-i. l. t.-l. t.-d. n. d.-l. s. b.-s. border-box et al., "A Novel Machine Learning Data Preprocessing Method for Enhancing Classification Algorithms Performance," ed. 16th EANN workshops: Proceedings of the 16th Engineering Applications of Neural Networks Conference WORKSHOPS: Association for Computing MachineryNew YorkNYUnited States, 2015, pp. 1-5.

[23] M. A. P. M. Md Manjurul Ahsan, Pritom Kumar Saha, Kishor Datta Gupta and Zahed Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," vol. 9, ed. technologies: MDPI, 2021.

[24] I. Pratama and A. E. Permanasari, Ardiyanto, I., & Indrayani, R., "A review of missing values handling methods on time-series data.," ed. In 2016 International Conference on Information Technology Systems and Innovation (ICITSI): IEEE, 2016, pp. 1-6.

[25] M. B. Kursa, & Rudnicki, W. R., "The all relevant feature selection using random forest.," ed. https://arxiv.org/abs/1106.5112: arxiv.org, 2011.

[26] N. M. K. S. A. Chinmay, "Optimization of the Random Forest Algorithm," ed. Advances in Data Science and Management, 2020, pp. 201-208.

[27] K. Grabczewski, & Jankowski, N., "Feature selection with decision tree criterion.," ed. In Fifth International Conference on Hybrid Intelligent Systems (HIS'05): IEEE, 2005, p. 6.

[28] B. T. J. a. A. M. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," vol. 2, ed. Journal of Applied Science And Technology Trends, 2021, pp. 20-28.

[29] D. B. Figueiredo Filho and J. A. S. Júnior, & Rocha, E. C., "What is R2 all about? ," vol. 3, ed. Leviathan: Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2011, pp. 60-68.

[30] B. F. Darst, Malecki, K. C., & Engelman, C. and D., "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data.," vol. 19, ed. BMC Genetics (BMC GENET): BioMed Central, 2018, pp. 1-6.

[31] Sebastian Raschka , a. Joshua Patterson , and C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence," ed: MDPI - Information, April, 2020.