A Review of the Data-Driven Prediction Method of Vehicle Fuel Consumption

Dengfeng Zhao 1, Haiyang Li 1,*, Junjian Hou 1, Pengliang Gong 2, Yudong Zhong 1, Wenbin He 1 and Zhijun Fu 1

1 Henan Provincial Key Laboratory of Intelligent Manufacturing of Mechanical Equipment, Mechanical and Electrical Engineering Institute, Zhengzhou University of Light Industry, Zhengzhou 45002, China; zhaodf@zzuli.edu.cn (D.Z.); houjunjian@zzuli.edu.cn (J.H.); zhongyd@zzuli.edu.cn (Y.Z.); hwb@zzuli.edu.cn (W.H.); fuzhijun2008@163.com (Z.F.)

2 Zhengzhou Senpeng Electronic Technology Co., Ltd., Zhengzhou 450052, China; gongpl@senptec.com

* Correspondence: haiyang0198@163.com

Abstract: Accurately and efficiently predicting the fuel consumption of vehicles is the key to improving their fuel economy. This paper provides a comprehensive review of data-driven fuel consumption prediction models. Firstly, by classifying and summarizing relevant data that affect fuel consumption, it was pointed out that commonly used data currently involve three aspects: vehicle performance, driving behavior, and driving environment. Then, from the model structure, the predictive energy and the characteristics of the traditional machine learning model (support vector machine, random forest), the neural network model (artificial neural network and deep neural network), and this paper point out that: (1) the prediction model of fuel consumption based on neural networks has a higher data processing ability, higher training speed, and stable prediction ability; (2) by combining the advantages of different models to build a hybrid model for fuel consumption prediction, the prediction accuracy of fuel consumption can be greatly improved; (3) when comparing the relevant indicts, both the neural network method and the hybrid model consistently exhibit a coefficient of determination above 0.90 and a root mean square error below 0.40. Finally, the summary and prospect analysis are given based on various models’ predictive performance and application status.

Keywords: fuel consumption; data-driven; machine learning; neural network; hybrid model

1. Introduction

With the development of the automotive industry, environmental pollution caused by vehicle exhaust emissions greatly impacts the human living environment and physical health [1]. The fuel consumed by vehicles mainly comes from non-renewable energy such as petroleum, and the problem of excessive fuel consumption and exhaust emissions urgently needs to be addressed. Accurately predicting a vehicle's fuel consumption can help drivers adjust their driving strategies, optimize fuel efficiency, save fuel costs, and reduce environmental pollution. Furthermore, by monitoring real-time fuel consumption data, potential accidents can be effectively prevented, thereby enhancing driving safety.

At present, the prediction methods for vehicle fuel consumption can be roughly classified into two categories (as shown in Figure 1): (1) a physical fuel consumption prediction model constructed based on the principle of vehicle dynamics; (2) a data-driven fuel consumption prediction model.
The first type of model is mainly built through mathematical formulas based on the internal structure of the vehicle and the working principles of components, such as the engine. The model’s transparency is high and can provide more accurate prediction results [2]. However, the research of this type of model is mainly focused on the fixed path in some specific areas, and the influence of different road types and weather conditions on fuel consumption is often ignored, resulting in a single data dimension and the poor applicability of models. For example, Chang et al. [3] used sensors installed on a certain road section to obtain vehicle state parameters at specific locations as model inputs, while Huang et al. [4] used traditional microscopic models (MOVES) to predict vehicle fuel consumption, only using data obtained under simple road conditions, resulting in an inability to guarantee the predictive effect of fuel consumption in practical applications.

The second type of model mainly relies on sensors and other on-board equipment to obtain a large number of vehicle operation data related to fuel consumption. By mining the optimal features from the data and establishing the nonlinear relationship between the data and fuel consumption, fuel consumption prediction is realized. Compared to the traditional physical model, the data-driven model is simple and easy to construct. It can automatically execute repetitive and tedious tasks, saving time and resources while ensuring good accuracy. The performance of various data-driven models is usually achieved through some evaluation indicators, including the coefficient of determination ($R^2$), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), scatter index (SI), and uncertainty with a 95% confidence level ($U_{95}$). The calculation equation and evaluation standard for each index are shown in Table 1.

When using physical models to predict fuel consumption, researchers are required to possess strong domain knowledge, and the application scope is limited, making it challenging to widely apply them to vehicles with different characteristics. The introduction of data-driven methods effectively addresses the limitations of physical models. Well-trained models in data-driven approaches can consider data from different factors as inputs, and their application extends to a wider range of vehicle types. This paper discusses the applications of traditional machine learning (Traditional ML) and the neural network method in the prediction models of fuel consumption based on data-driven models (the process of realizing a fuel consumption prediction by the model is shown in Figure 2). Therefore, based on the existing studies listed in this paper, the above two commonly used data-driven methods are summarized from the two aspects of research years and the number of researchers (as shown in Figure 3).
Table 1. The calculation equation and evaluation standard for each index.

<table>
<thead>
<tr>
<th>Index</th>
<th>Equation</th>
<th>Evaluation Standard</th>
<th>Equation No.</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>$R^2 = 1 - \frac{\sum_{m=1}^{n} (y_m - y'<em>m)^2}{\sum</em>{m=1}^{n} (y_m - \bar{y})^2}$</td>
<td>Range: 0 to 1 Higher is best</td>
<td>(1)</td>
<td>[5–16]</td>
</tr>
<tr>
<td>MSE</td>
<td>$MSE = \frac{1}{n} \sum_{m=1}^{n} (y_m - y_m)$</td>
<td>Lower is best</td>
<td>(2)</td>
<td>[17–20]</td>
</tr>
<tr>
<td>RMSE</td>
<td>$RMSE = \sqrt{\frac{1}{n} \sum_{m=1}^{n} (y_m - y_m)}$</td>
<td>Lower is best</td>
<td>(3)</td>
<td>[8–11,13,14,21–24]</td>
</tr>
<tr>
<td>MAE</td>
<td>$MAE = \frac{1}{n} \sum_{m=1}^{n}</td>
<td>y_m - y_m</td>
<td>\times 100%$</td>
<td>Lower is best</td>
</tr>
<tr>
<td>MAPE</td>
<td>$MAPE = \frac{1}{n} \sum_{m=1}^{n} \frac{</td>
<td>y_m - y_m</td>
<td>}{y_m} \times 100%$</td>
<td>Lower is best</td>
</tr>
<tr>
<td>SI</td>
<td>$SI = \frac{RMSE}{y}$</td>
<td>$SI &lt; 0.05$: higher accuracy; $0.05 &lt; SI &lt; 0.1$: Good accuracy</td>
<td>(6)</td>
<td>[28,29]</td>
</tr>
<tr>
<td>$U_{95}$</td>
<td>$U_{95} = 1.96\sqrt{(SD^2 + RMSE^2)}$</td>
<td>Lower is best</td>
<td>(7)</td>
<td>[28,29]</td>
</tr>
</tbody>
</table>

where $y_m$ is the $m$ predicted fuel consumption in the test sample, L/h; $y_m$ is the true fuel consumption of the $m$ in the test sample, L/h; $n$ is the number of samples; $y$ is the average of the true fuel consumption, L/h; $SD$ is the standard deviation of the variance between true and predicted values.

Figure 2. The general process of using models to predict fuel consumption.

Figure 3. Research status of data-driven fuel consumption prediction methods.

Machine learning is a classic data-driven fuel consumption prediction method [30] that includes support vector machine (SVM), random forest (RF), decision tree (DL), etc. For
example, Heni et al. [31] used SVM and gradient boosting machines to perform nonlinear regression analysis on data, and a large number of experiments based on real conditions have demonstrated the superiority of machine learning methods in fuel consumption prediction models. Hamed et al. [5] established a functional relationship between vehicle speed and fuel consumption based on a support vector machine, with a $R^2$ of up to 0.97. Abukhalil et al. [32] used on-board diagnostic (OBD-II) system data to construct an SVM model to estimate fuel consumption, with an RMSE of 2.43. SVM has a concise and interpretable structure, but it is more suitable for data processing on a smaller scale. On the other hand, the random forest algorithm can handle high-dimensional data and perform well in predicting fuel consumption [33] by effectively reducing dimensionality. For example, Gong et al. [34] considered 21 factors that cause fuel consumption to establish a random forest fuel consumption model with a predictive accuracy of 86%. Zhang et al. [35] established a fuel consumption estimation model based on the least squares method using vehicle speed and acceleration. Zhu et al. [36] proposed a prediction model based on the improved C4.5 decision tree and verified the effectiveness of the model by relying on a set of test data under the expressway scenario. In addition, the application of gradient boosting algorithms [37,38], LightGBM [39], and linear regression (LR) [40] in fuel consumption prediction models has also achieved good results. In order to give full play to the advantages of traditional machine learning methods, Li [28] and Mahzad [41] et al. developed multiple hybrid models, including the Aquila optimizer and extreme gradient boosting (AO-XGB), black widow optimization algorithm and extreme gradient boosting (BWOA-XGB), AO-SVM, AO-RF, etc. The results show that such models have better generalization ability and more reliable prediction accuracy.

Traditional machine learning methods have limited data processing capabilities and struggle to adapt to complex large-scale datasets. As a method of solving complex problems in the field of engineering, neural networks are more likely to capture minor changes in complex data, to have good generalization capabilities, and to adapt to changes in new data and new environments. At present, this technology has been widely used in the predictions of emissions and fuel consumption of vehicles [42–44], ships [45–47], and aircraft [48,49]. Katreddi et al. [50] established an artificial neural network model using engine speed, vehicle speed, and other data to predict the fuel consumption of a vehicle in a single journey. The results show that the performance of the artificial neural network (ANN) is better than traditional machine learning methods, such as LR and RF. Zargarnezhad et al. [51] successfully estimated the additional fuel consumption caused by an increase in vehicle weight using an artificial neural network based on the relationship between changes in vehicle weight and engine displacement and fuel consumption, with an $MSE$ of 0.308. Topić et al. [6] trained different fuel consumption prediction models by the driving cycle data obtained by GPS and CAN bus and proved that the ANN model has higher accuracy and a higher ability of execution. Neural networks represented by ANN typically require data preprocessing and lack memory functionality during the prediction process. With the development of digital technology and the enrichment of network storage functions, deep neural network learning methods represented by a recurrent neural network (RNN), convolutional neural network (CNN), and multi-layer perceptron neural network (MLP) can directly extract features from original data without relying on feature engineering, reducing the time cost of human participation in feature extraction. For example, Ali et al. [52] used the multi-layer perceptron (MLP) neural network to build a fuel consumption model with the total weight and vehicle speed of trucks as input and obtained a relatively low $MSE$ of 0.0017. Due to the influence of time series on vehicle fuel consumption, conventional neural network methods are ineffective in addressing this issue. Panapakidis et al. [53] proposed a mixed model based on RNN, which fully evaluated the influence of exogenous parameters on ship fuel consumption and proved the superiority of the RNN prediction model. RNN is proficient at extracting vehicle fuel consumption features related to time series, especially with long and short-term memory (LSTM) networks, which can handle data with long time intervals and significant delays.
more effectively. Bougiouklis et al. [54] proposed an electric vehicle energy management strategy based on the LSTM neural network, which reduced energy consumption by 24.03%. Based on the advantages of CNN in image processing, Valido et al. [55] captured the image sequence of vehicles running on the road through the camera sensor, then located the position coordinates of each vehicle in the image through CNN, and finally realized the estimation of vehicle emission and fuel consumption based on distance information and speed, with an average error of 5.48%. In summary, the neural network method has good fault tolerance and robustness, can deal with missing data and noise, and can adjust model parameters adaptively. Therefore, applying it in the prediction models of fuel consumption can obtain prediction results that are highly correlated with the actual fuel consumption of vehicles.

In this paper, various data-driven prediction models of fuel consumption in recent years were reviewed from the aspects of model construction, data acquisition, and prediction performance. By comparing the traditional machine learning method with the neural network method, the paper points out that: (1) the fuel consumption prediction model based on the neural network can better mine the feature information related to fuel consumption in the data, can establish the nonlinear relationship between sensor data and fuel consumption prediction, and has better model generalization ability, higher stability, and better prediction accuracy; (2) prediction models with a single method or a single dimension tend to pay too much attention to the details of the training data set and cannot be generalized, which easily leads to the overfitting of the model and ignores the impact of the interaction between different factors on fuel consumption; (3) the application of the hybrid model and multivariate data fusion technology can fully consider the influence of multi-dimensional factors such as person-vehicle-road on fuel consumption, and the research in this field will be the future development direction.

The remaining structure of this paper is as follows: The second section categorizes and summarizes the relevant data that affect fuel consumption and introduces the methods of obtaining different types of data. The third section discusses and analyzes fuel consumption prediction methods’ characteristics and research status. Finally, the paper summarizes the characteristics of different fuel consumption prediction models and gives prospects.

2. Data Analysis of Vehicle Fuel Consumption

The data required by the prediction model of fuel consumption mainly depend on variables that have an impact on vehicle fuel consumption. Generally speaking, they can be roughly divided into three categories: vehicle inherent variables, driving behavior variables, and driving environment variables (as shown in Table 2). Vehicle inherent variables include vehicle and engine model, engine capacity, total vehicle mass, etc., which can be obtained according to the information resources provided by vehicle manufacturers. Driving behavior variables are mainly based on the changes in vehicle running state data caused by the driver’s behaviors such as stepping on the pedal and turning during vehicle starting and stopping and vehicle running. Driving environment variables mainly come from the influence of uncontrollable factors such as weather factors, altitudes, and the road slope on the vehicle’s running state. These two types of data mainly include driving speed and acceleration, engine speed and torque, load rate of engine and revolution, driving distance, etc. The changes in vehicle operating data caused by the above variables can be obtained through onboard sensing devices such as GPS, gyroscope, OBD-II, and an on-board controller area network (CAN). Considering the convenience of data uploading and the portability of devices, smartphone sensors and apps can also be used to obtain information such as vehicle operation data, driving behavior, and the real fuel consumption of vehicles. Research manuscripts reporting large datasets that are deposited in a publicly available database should specify where the data have been deposited and provide the relevant accession numbers. If the accession numbers have not yet been obtained at the time of submission, it will state that they will be provided during review. They must be provided before publication.
Table 2. Classification of fuel consumption data.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Relevant Data</th>
<th>Collection Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle inherent variables</td>
<td>Vehicle and engine model, engine capacity, total vehicle mass</td>
<td>Provided by the vehicle manufacturer</td>
</tr>
<tr>
<td></td>
<td>Driving speed and acceleration, engine speed and torque, engine load rate and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>revolution, driving distance, the real fuel consumption of vehicles</td>
<td></td>
</tr>
<tr>
<td>Driving behavior variables</td>
<td>Weather factors, altitude, road slope, and other road conditions information</td>
<td>GPS, Radar, Infrared ray</td>
</tr>
<tr>
<td>Driving environment variables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Due to the possibility of errors or communication failures during the data collection process, it is necessary to filter the original data to eliminate invalid and redundant data, including duplicate data, extreme data, and not-running vehicle data (such as data with a speed of 0 or GPS information not updated for a long time). For the filtered effective data, depending on feature engineering, is necessary to establish a dataset with a high correlation to fuel consumption, aiming to improve the robustness and convergence speed of the prediction model (as shown in Figure 4). This process mainly involves two stages: feature dimensionality reduction and feature selection. Principal component analysis (PCA) and Pearson correlation coefficient (PCC) are the most commonly used techniques. PCA is commonly used for unsupervised dimensionality reduction tasks. It employs linear transformations to map the original features to a new low-dimensional space. By maximizing the variance, PCA enhances the separability between samples, thereby preserving as much original data information related to fuel consumption as possible. The PCC method is used to measure the strength of the linear relationship between two variables, and a correlation coefficient between −1 and 1 is obtained by calculating the variance between variables to achieve the best variable screening. In addition, sensitivity analysis is also used to determine the influence of an input variable on an output variable. The higher the sensitivity coefficient, the stronger the correlation between the variables (see Equation (8)):

\[
W_i = \frac{MSE_i}{MSE}
\]

where \(W_i\) is the sensitivity coefficient of the model to the variable \(i\); \(MSE_i\) is the MSE of the model without the variable \(i\).

Figure 4. The general process of dataset creation.

3. Prediction Models of Vehicle Fuel Consumption

Machine learning is a method that uses algorithms and statistical models to learn general rules from observed data to predict unknown variables, including traditional machine learning (RF, SVM, LR, etc.) and neural network learning (ANN, FNN, BPNN, LSTM, etc.). It has the advantages of a simple structure, high prediction accuracy, and strong
ability in data recognition, which can facilitate the establishment of nonlinear relationships between observation data and automotive fuel consumption and can be applied more in the field of automotive fuel consumption prediction. This part mainly introduces the structure, principle, and performance of fuel consumption prediction models.

3.1. SVM Model

SVM is mainly used to solve binary classification tasks and relies on the kernel function to realize mapping from low-dimensional data to high-dimensional data. Its working principle is shown in Figure 5. Figure 6 shows the process of using support vector machines for fuel consumption prediction.

Figure 5. The basic principle of SVM.

Figure 6. Fuel consumption forecasting process based on SVM.

SVM is suitable for processing small sample data and can select different kernel functions based on the type of datasets, which can be well applied to the fitting problem of fuel consumption models [56,57]. Wang et al. [58] used SVM to predict aircraft fuel consumption, with an average estimation error of $-0.039\%$. Hussain et al. [7] used onboard sensor data to predict bus fuel consumption, and the $R^2$ of the SVM model was 0.95. Araújo et al. [21] used the SVM algorithm to predict the influence of road–tire interaction on vehicle energy consumption, and the results showed that using more data as input could improve the prediction performance of the model and reduce the error. However, the aforementioned studies used partially abstract input data, which posed challenges in terms of measurability. In response, Liu et al. [25] simplified the data collection process by using the most common variables, such as engine speed, vehicle speed, and acceleration, as inputs. They also achieved favorable prediction results, with a mean absolute error (MAE) of less than 0.16. Support vector regression (SVR) is a special algorithm for solving regression problems in SVM [59].

Nevertheless, studies that employed a single SVM model to predict fuel consumption cannot guarantee idealized prediction results. Ma et al. [60] used the Asymmetric ε-band fuzzy support vector regression based on the data domain description (ASVDD) to predict fuel consumption, and the MSE was less than 0.00127. Using a multi-fusion algorithm model to forecast fuel consumption can not only improve the overall performance of the model but can also improve the overfitting problem caused by noise and other factors. For example, Li et al. [61] proposed a coupled model of extreme learning machine and support vector
machine (SVM-ELM) to predict energy consumption, with an \( R^2 \) of 0.99. In addition, when using the least squares method (LSM) [62] and genetic algorithm (GA) [17] in the support vector machine model, the prediction performance can also be significantly improved.

According to the input of different data sources and data characteristics, the SVM model has different adaptability and performance. The results of relevant studies are shown in Table 3.

### Table 3. Analysis of fuel consumption models based on SVM.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Model</th>
<th>Inputs</th>
<th>Vehicle Type</th>
<th>Performance</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abukhalil et al. [32]</td>
<td>SVM</td>
<td>Engine speed, revolutions per minute, speed, etc.</td>
<td>Passenger vehicles</td>
<td>RMSE: 2.43</td>
<td>Processing large-scale data is inefficient and time-consuming; the model accuracy is low when dealing with noisy data sets. The SVM model is relatively simple and has low requirements for hardware and software.</td>
</tr>
<tr>
<td>Zeng et al. [63]</td>
<td>SVM</td>
<td>Data collected from GPS and CAN bus (trip distance, speed, etc.)</td>
<td>Probe vehicles</td>
<td>( R^2 ): 0.92</td>
<td></td>
</tr>
<tr>
<td>Hussain et al. [7]</td>
<td>SVM</td>
<td>Information from On-board Sensors and records (traveled distance, hour of the day, driver ID, bus ID, etc.)</td>
<td>City buses</td>
<td>( R^2 ): 0.95</td>
<td></td>
</tr>
<tr>
<td>Capraz et al. [8]</td>
<td>SVM</td>
<td>Trip distance, speed, vehicle weight, acceleration, road slope, etc.</td>
<td>Passenger vehicles</td>
<td>( R^2 ): 0.94</td>
<td></td>
</tr>
<tr>
<td>Arajjo et al. [21]</td>
<td>SVM</td>
<td>Load, speed, pendulum test value, mean texture depth, estimating the surface texture depth; load, speed, mean texture depth</td>
<td>Various types of vehicles</td>
<td>RMSE: 0.303</td>
<td></td>
</tr>
<tr>
<td>Liu et al. [25]</td>
<td>SVR</td>
<td>Engine speed and torque, speed, acceleration, engine oil temperature, etc.</td>
<td>Passenger vehicles</td>
<td>MAE: &lt;0.16</td>
<td>The performance is highly dependent on the model fit and the characteristics of the data set. High computing resources and storage space are required.</td>
</tr>
<tr>
<td>Ahmadi et al. [22]</td>
<td>LSSVM</td>
<td>Rotation speed, temperature of heat source, pressure, etc.</td>
<td>Stirling engine</td>
<td>RMSE: 0.067 ( R^2 ): 0.98</td>
<td></td>
</tr>
<tr>
<td>Wang et al. [17]</td>
<td>GA-SVM</td>
<td>Data from sensors (engine speed and torque, temperature, air mass flow rate, etc.)</td>
<td>Light diesel engine</td>
<td>MSE: 1.344 ( R^2 ): 0.967</td>
<td></td>
</tr>
<tr>
<td>Li et al. [64]</td>
<td>NSGA-SVM</td>
<td>Speed, load, engine speed, cylinder pressure, etc.</td>
<td>Gasoline engine</td>
<td>Relative error: &lt;3%</td>
<td></td>
</tr>
</tbody>
</table>

#### 3.2. RF Model

The RF model is a supervised learning algorithm that uses an optimal decision tree to learn rules from a given sample. The importance of each data can be estimated during the execution of the algorithm, and missing values can be automatically processed. The working principle is shown in Figure 7. In Figure 8, bagging stage 1 performs random sampling with dropout to achieve the reorganization of the new dataset; in bagging stage 2, the decision tree optimizes the relevant data and extracts features; the voting period is responsible for voting on the results generated by each decision tree to determine the optimal fuel consumption output.

![Figure 7. The basic principle of RF.](image-url)
RF has fewer parameters, which can compensate for the difficulty of parameter tuning in SVM and exhibit better performance. This point was demonstrated in the research conducted by Pereira et al. [65]. Therefore, RF is widely applied in the prediction of fuel consumption [66] and energy consumption [67]. Perrotta et al. [10] established three truck fuel consumption prediction models, namely SVM, ANN, and RF, with determination coefficients of 0.83, 0.85, and 0.87, respectively. Due to the mentioned studies’ specific focus on articulated trucks and the limited consideration of variables during the data acquisition phase, such as neglecting the impact of factors such as temperature and driving behavior on fuel consumption [68–70], the model’s accuracy was compromised, resulting in poorer precision. In the literature [71], Massoud et al. used RF to analyze the relationship between driving behavior data and fuel consumption and took characteristic parameters representing speed and engine speed as input, $R^2$ and $MSE$ are 0.896 and 1.506. Yang et al. [18] considered the influence of vehicle factors, environmental factors, and driving behavior factors on fuel consumption, and the $MAE$ of the random forest fuel consumption prediction model was 0.63, which greatly improved the generalization level of the model.

When considering the influence of multiple factors as inputs, the random forest (RF) model exhibited poorer performance. Therefore, Hu et al. [72] proposed a hybrid model consisting of RF, XGBoosting, and multiple linear regression (MLR) methods to predict ship fuel consumption. Compared to using a single model, this hybrid model achieved smaller error values. Shi et al. [29] integrated the improved arithmetic optimization algorithm with RF, and the application effect of the model was significantly improved (considering $R^2$ and RMSE).

The prediction model of fuel consumption established based on RF can effectively identify the nonlinear relationship between different variables and provide prediction results that are highly correlated with actual fuel consumption, especially the application of the coupling model, which can effectively reduce the error of the fuel consumption prediction model. Table 4 summarizes the literature that used the RF model to predict fuel consumption.
Table 4. Result analysis of fuel consumption models based on RF.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Model</th>
<th>Inputs</th>
<th>Vehicle Type</th>
<th>Performance</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hassan et al. [9]</td>
<td>RF</td>
<td>Vehicle speed, vehicle specific power, engine speed, engine stress</td>
<td>Passenger vehicles</td>
<td>RMSE: 0.15</td>
<td>The effect of processing high-dimensional sparse data is not good. It is difficult to explain the specific reasons for the model prediction; when the number of trees is large, high computing resources and storage space are required.</td>
</tr>
<tr>
<td>Gong et al. [34]</td>
<td>RF</td>
<td>21 variables extracted from the driver-vehicle-road-environment</td>
<td>Heavy-duty diesel trucks</td>
<td>Accuracy: 86.58%</td>
<td></td>
</tr>
<tr>
<td>Perrotta et al. [10]</td>
<td>RF</td>
<td>Vehicle speed, acceleration, the torque and revolutions of the engine, etc.</td>
<td>Trucks</td>
<td>RMSE: 4.64  R²: 0.87</td>
<td></td>
</tr>
<tr>
<td>Yu et al. [11]</td>
<td>RF</td>
<td>Mileage and speed, temperature, air pressure, etc.</td>
<td>Trucks</td>
<td>RMSE: 5.073  R²: 0.98</td>
<td></td>
</tr>
<tr>
<td>Yao et al. [23]</td>
<td>RF</td>
<td>Driving data collected from smartphone applications (speed, acceleration, etc.)</td>
<td>Taxis</td>
<td>RMSE: 0.783</td>
<td></td>
</tr>
<tr>
<td>Yang et al. [18]</td>
<td>RF</td>
<td>Engine power and the number of cylinders, driving speed, driving habits, temperature, wind speed, etc.</td>
<td>Light-duty vehicles</td>
<td>MAE: 0.63  MSE: 0.805</td>
<td></td>
</tr>
<tr>
<td>Yang et al. [24]</td>
<td>RF</td>
<td>Engine speed, engine torque, speed, acceleration, deceleration, etc.</td>
<td>Gain combine harvesters</td>
<td>MAE: 0.24  RMSE: 0.14</td>
<td></td>
</tr>
<tr>
<td>Yu et al. [12]</td>
<td>Hybrid model</td>
<td>Mileage and speed, temperature, air pressure, etc.</td>
<td>Long-distance vehicles</td>
<td>R²: 0.976</td>
<td>The model is complex and requires high hardware and software.</td>
</tr>
</tbody>
</table>

3.3. Neural Network Model

As a mathematical model that mimics the structure and function of the biological brain, ANN can accept multiple data inputs to achieve single- or multiple-result outputs, and this is something SVM and RF models do not possess. Figure 9 shows the structure of the artificial neural network model.

![Figure 9. The basic structure of the artificial neural network.](image)

The neural network has good nonlinear mapping ability [73], which can automatically identify and learn the characteristic information of input data [74,75]. Moreover, it has a strong capability in parallel computing and can show high efficiency in processing large-scale data [76,77]. Huang et al. [78] used a radial basis function neural network (RBFNN) to predict fuel consumption with an accuracy of 85%. However, it should be noted that the model they proposed was trained only on a local dataset, which implies its strong regional bias and limited applicability in other regions. As a result, Wysocki et al. [13] trained a fuel consumption model using heavy-duty truck driving data collected over a five-year span. They employed an ANN model and achieved an RMSE of 0.32. However, it is important to note that historical data can be influenced by environmental changes and vehicle performance degradation, leading to unreliable predictions. To address the issue of low data reliability, Ling et al. [79] proposed a model predictive control (MPC) framework based on artificial neural networks. This framework relied on real-time predictions of
vehicle speed to effectively tackle the problem. Simulation experiments demonstrated that HDVs (heavy-duty vehicles) controlled by MPC achieved a 5.9% reduction in fuel consumption. Asher [26] and Sun [14] et al. applied an artificial neural network to the fuel consumption prediction of hybrid electric vehicles, and the MAE of the model was between 0 and 0.1%. In addition, there are a few studies on the fuel consumption prediction model based on a feedforward neural network (FNN) [80]. Topić et al. [6] used vehicle speed to predict the fuel consumption of a bus, and the $R^2$ of FNN model was more than 0.97.

There are interactions among different variables that display clear non-linear relationships. However, conventional ANN can only handle variables with linear variations. In contrast, the backpropagation neural network (BPNN) trains and adjusts the weights of connections in the network using nonlinear differential functions and is capable of effectively addressing complex nonlinear problems [81]. Du et al. [82] proposed a BPNN model with a structure of 9-10-1, analyzed fuel consumption levels from two dimensions of time and space, and comprehensively described the relationship between fuel consumption and various influencing factors, and the accuracy of the model reached 81.7%.

While the study considered multiple factors influencing fuel consumption, the accuracy of using a standalone BP neural network for fuel consumption prediction is limited. Therefore, some studies have coupled multiple models to predict fuel consumption [83,84]. For example, Shang et al. [19] used data such as vehicle speed and GPS coordinates, combined with the hidden Markov model (HMM) and BPNN to predict fuel consumption, and the MSE of the model was less than 0.06 and the $R^2$ was more than 0.95. Similar coupling models for fuel consumption prediction also include the BPNN model based on the genetic annealing algorithm (GSA) and the BPNN model based on the Cauchy multi-verse optimizer (CMVO).

The results of fuel consumption prediction models based on different neural networks are shown in Table 5. Compared to the single neural network prediction model, the prediction accuracy of fuel consumption can be further improved by coupling the neural network with other methods.

Table 5. Result analysis of fuel consumption models based on NN.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Model</th>
<th>Inputs</th>
<th>Vehicle Type</th>
<th>Performance</th>
<th>Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wysocki et al. [13]</td>
<td>ANN</td>
<td>Data from the CAN bus (engine speed and torque, etc.)</td>
<td>Heavy-duty trucks</td>
<td>RMSE: 0.32</td>
<td>$R^2$: 0.99</td>
</tr>
<tr>
<td>Witaszek [85]</td>
<td>ANN</td>
<td>Vehicle speed and acceleration, road slope, throttle opening degree, selected gear number, and engine speed</td>
<td>Passenger vehicles</td>
<td>Relative error: &lt;3%</td>
<td></td>
</tr>
<tr>
<td>Soofastaei et al. [86]</td>
<td>ANN</td>
<td>Data from past records (payload, total resistance, actual speed)</td>
<td>Haul trucks</td>
<td>$R^2$: 0.903</td>
<td></td>
</tr>
<tr>
<td>Asher et al. [26]</td>
<td>ANN</td>
<td>Data from OBD-II (speed, acceleration, engine speed, etc.)</td>
<td>Hybrid vehicles</td>
<td>MAE: 0–0.1%</td>
<td></td>
</tr>
<tr>
<td>Schone et al. [87]</td>
<td>FNN</td>
<td>Distance, seven predictors derived from vehicle speed and road slope</td>
<td>Heavy-duty vehicles</td>
<td>RMSE: 0.0132</td>
<td>$R^2$: 0.91</td>
</tr>
<tr>
<td>Topić et al. [6]</td>
<td>FNN</td>
<td>Speed, acceleration, slope time series</td>
<td>City buses</td>
<td>$R^2$: &gt;0.97</td>
<td></td>
</tr>
<tr>
<td>Du et al. [82]</td>
<td>BP</td>
<td>Data from records (time, location, speed, road condition, driver’s personal information, etc.)</td>
<td>Various types of vehicles</td>
<td>Accuracy: 81.7%</td>
<td>Nonlinear mapping strength; easy to fall into local optimal solution; susceptible to initial values; when the data scale is large, the device requires high configuration.</td>
</tr>
<tr>
<td>Zhao et al. [88]</td>
<td>BP</td>
<td>Data from OBD-II and GPS (distance, acceleration, speed, etc.)</td>
<td>Taxis</td>
<td>Accuracy: 92.46%</td>
<td></td>
</tr>
<tr>
<td>Shang et al. [19]</td>
<td>HMM-BP</td>
<td>Data from records (vehicle ID, vehicle speed, moving direction, GPS longitude, latitude, etc.)</td>
<td>Taxis</td>
<td>MSE: &lt;0.06</td>
<td>$R^2$: &gt;0.95</td>
</tr>
<tr>
<td>Zhou et al. [89]</td>
<td>GSA-BP</td>
<td>Engine speed and torque, vehicle speed, load rate, driving distance, etc.</td>
<td>Heavy duty vehicles</td>
<td>Accuracy: 96.51%</td>
<td>The performance is highly dependent on the model fit and the characteristics of the data set. High computing resources and storage space are required.</td>
</tr>
<tr>
<td>Chen et al. [90]</td>
<td>CMVO-BP</td>
<td>Vehicle speed, engine speed, and torque</td>
<td>Truck</td>
<td>Accuracy: 97.5%</td>
<td></td>
</tr>
</tbody>
</table>
3.4. Deep Neural Network

The neural network model represented by ANN and FNN is a static prediction process based on known historical data. However, the actual vehicle operation data are constantly changing, which may lead to a large deviation between the predicted value and the actual value [91]. Deep neural networks (DNN) are composed of neurons at multiple levels [92]. The automatic learning of data and automatic feature extraction can be realized through multiple feedback training, which is suitable for establishing the dynamic variation process between feature data and fuel consumption [93]. Li et al. [94] trained a well-performing MLP fuel consumption model based on factors such as climate conditions and vehicle characteristics to obtain data. Ziolkowski et al. [27] applied the Pearson correlation coefficient method to an MLP model and used 1750 passenger vehicle data to construct a friendly model with a structure of 22-10-3. They were able to effectively control the MAPE within the range of 5% to 8%. While MLP models can address the negative impact of noise factors, they lack the ability to perform recurrent processing compared to RNN and CNN models. MLP models can only propagate information in a forward direction and do not have the capability to autonomously handle relevant data features. RNN is a kind of neural network with time delay characteristics and contains a loop structure inside, which enables it to recursively propagate the input and retain the previously useful data processing state [95]. For example, Xu et al. [20] used a generalized recurrent neural network (GRNN) to establish a relationship model between driving behavior and fuel consumption and took the speed obtained based on different routes as input to obtain lower relative error and MSE. Kanarachos et al. [96] selected RNN based on the nonlinear AutoRegressive with ExogenousInputs Model (NARX) to predict the instantaneous fuel consumption of the vehicle, and the error between the prediction result and actual fuel consumption was less than 6%. The above two studies relied on the Internet of Vehicles and smartphones for data collection, respectively. However, the data upload process is susceptible to network delays, and the RNNs with general network structures struggle to handle the long-term dependencies in the data and are prone to issues, such as gradient vanishing and gradient exploding. LSTM adds a gating mechanism based on RNN, which can effectively solve the problem of long sequence information attenuation and better deal with long lag data. The structure of LSTM is shown in Figure 10.

![Figure 10. The cycle cell structure of LSTM.](image)

Based on the advantages of LSTM in fuel consumption prediction, Ping et al. [97] established an LSTM neural network with a different number of hidden nodes combined with driving behavior and traffic condition data to predict fuel consumption, and the prediction accuracy could reach 84.7%. Kan et al. [98] proposed a heavy truck fuel consumption estimation model based on LSTM, with an average error of 0.137. Jain et al. [99] used LSTM to monitor the instantaneous fuel economy of vehicles, and the overall accuracy of the model exceeded 98%. Based on the data obtained in various scenarios, Wang et al. [100]
used LSTM to predict vehicle fuel consumption, and the error was less than 0.1. With a large number of weights for LSTM, it takes several iterations to obtain a well-trained model, and overfitting is prone to occur when the number of input data is insufficient. Therefore, Hua et al. [101] adopted model pre-training and transfer learning to achieve the high-level prediction of energy consumption.

CNN uses convolution operation to extract features from input data, which has a local connection and power-sharing characteristics. Compared to RNN, CNN is suitable for processing data with spatial structures. CNN is composed of a convolutional layer, a pooling layer, and a fully connected layer, as shown in Figure 11. The convolution layer performs convolution calculations for input and extracts features; the pooling layer reduces the computation by reducing the dimension of data; and the full-connection layer is responsible for the output of results.

![Figure 11. The basic structure of CNN.](image)

CNN has certain advantages in image processing [102,103] but is rarely used in vehicle fuel consumption prediction. Hien et al. [15] used a one-dimensional convolutional neural network to estimate the total fuel consumption of vehicles on highways and urban roads, and the $R^2$ was 0.99. Yan et al. [104] applied CNN to the energy management of hybrid electric vehicles, effectively improving the fuel economy of vehicles. Han et al. [105] proposed a coupling model concept based on CNN, and Metlek [16] verified the coupling model of CNN and LSTM with 13 different input parameters and obtained a high $R^2$ of 0.974.

Deep neural networks have strong adaptability and learning ability, which can accept direct input from original data and transform it into more abstract representations to learn more complex functions from the data. Deep neural networks are known for their strong dependence on input data. The more data that are available for training, the better the performance of the model tends to be. This is particularly true when compared to fuel consumption prediction models built using other methods. Although deep neural networks can approximate any nonlinear continuous function with arbitrary accuracy, it cannot explain the complex decision-making process, and the model visualization is not strong; DNN has higher requirements for data quality, which increases the cost of manual annotation. In addition, the high complexity of the deep neural network model requires higher hardware performance of computing equipment.

4. Summaries and Prospects

4.1. Summaries

This paper mainly reviewed data-driven fuel consumption forecasting methods, including SVM, RF, Ann, BP, and RNN. By comparative analysis, the advantages and disadvantages of various fuel consumption prediction methods were summarized (Table 6), the prediction results were compared (based on the two evaluation indicators of $R^2$ and RMSE, as shown in Figure 12), and the following conclusions were drawn:
(1) In the study of the data-driven fuel consumption prediction models, since the fuel consumption process of vehicles is affected by multiple time-varying factors (such as the vehicle running state, driver habits, and driving environment), it is necessary to further consider the problem of poor fit caused by data coupling and so on. To solve this problem, PCA and other methods can be used to reduce the extraction of redundant features and solve the problem of poor model performance on high-dimensional data sets; the Pearson correlation coefficient method can also be used to analyze and screen out features highly correlated with fuel consumption as the input of the model to further ensure that the model has sufficient accuracy.

(2) Traditional machine learning methods have good predictive performance, but some methods need to extract features manually. Existing studies mainly concentrate on the use of a single scenario set, and the model has poor applicability and limited promotion. Therefore, in the data collection stage, considering the fusion of multi-dimensional features for fuel consumption modeling can effectively improve the accuracy and enhance the generalization capacity of the model.

(3) The prediction models of fuel consumption based on neural networks have high accuracy and stability in prediction, but they are too dependent on the size of input data. When the input data are insufficient, it is easy to show poor generalization ability or overfitting problems. To solve this problem, data enhancement can be used to increase the number of samples and maximize the utilization of sample data.

(4) The accuracy of fuel consumption prediction models largely depends on the quality and quantity of input data. Vehicle sensor data are widely used for their advantages of accuracy, reliability, large data volume, and low cost, but there are some problems, such as transmission delay. Using smartphones to obtain data is more real-time, efficient, and convenient. Therefore, in the future, rapid and comprehensive data collection can be achieved by combining onboard sensing devices and smartphones. In addition, when using large-scale datasets for model training, the generalization ability of the model can be effectively improved by using normalization and other processing methods in the preprocessing stage.

(5) The hybrid fuel consumption model is composed of different machine learning methods, which can synthesize the advantages of multiple models to deal with more complex tasks, with strong nonlinear expression ability and good model robustness. However, the structure of this model is complex, the calculation is large, the parameters are not easy to determine, and there are drawbacks in practical application.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretability</td>
<td>Good</td>
<td>Middle</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>Efficiency</td>
<td>High</td>
<td>Middle</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Low</td>
<td>Middle</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Advantages</td>
<td>Simple model, suitable for processing small sample data (SVM) and high dimensional data (RF)</td>
<td>Strong nonlinear mapping ability, relatively simple structure</td>
<td>Can process data related to timing (RNN, LSTM), automatic feature extraction, model stability</td>
<td>Suitable for all kinds of scenarios, accepts input from different types of data</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Features need to be extracted manually, poor performance when dealing with large amounts of data</td>
<td>Easy to fall into local optimality (BPNN), features related to timing cannot be obtained</td>
<td>Computationally heavy, over-reliance on the amount of input data</td>
<td>The model structure is complex, and parameter adjustment difficulty</td>
</tr>
</tbody>
</table>
(a) \quad \text{(b)}

Figure 12. (a) Comparison of results based on $R^2$; (b) comparison of results based on RMSE.

From the figure above, the following conclusions can be drawn: (1) Neural network methods, such as BPNN and DNN, applied in fuel consumption prediction models, can provide relatively accurate prediction results. (2) Hybrid prediction models combining machine learning and neural network techniques leverage the advantages of different models, leading to highly correlated predictions with actual fuel consumption values. (3) Based on the error analysis results, deep neural networks and hybrid models demonstrate the best performance. The overall trend of the graphs is stable, and the error levels are relatively consistent.

In summary, the current research on fuel consumption prediction faces challenges in simultaneously meeting the requirements of predictive performance regarding input data and model selection. On the one hand, some researchers tend to overly consider multiple factors influencing fuel consumption, which increases the difficulty in data processing and leads to inadequate model accuracy. On the other hand, some studies lean towards selecting simpler prediction models and a smaller, more manageable dataset to achieve higher predictive performance. In conclusion, by using a rational dataset and selecting suitable neural network methods or hybrid models, it is possible to obtain satisfactory fuel consumption prediction results. However, this approach may increase the complexity of the research, but it can yield stable and reliable outcomes.

4.2. Prospects

The application of traditional machine learning methods, neural network methods, and hybrid models has broken the shackle of traditional fuel consumption prediction methods, and more information related to fuel consumption can be mined through big data and other technical means to accurately predict fuel consumption levels. The neural network model can map the nonlinear relationship between input and output well, and the prediction performance is more stable; the hybrid model can play the advantages of different prediction models more comprehensively, can balance the disadvantages of different methods, and has high prediction accuracy. Therefore, establishing the relationship between on-board data and fuel consumption prediction by coupling different neural network methods will be the development trend in the field of vehicle fuel consumption prediction.

The development of automobile intelligence, data acquisition, and processing makes them more intelligent, efficient, safe, and reliable. Considering the influence of drivers’ driving style, vehicle performance, road environment, and other factors on fuel consumption, it is also the development trend of this field to apply driver-vehicle-road related multivariate data coupling into the input of fuel consumption prediction models to establish a more comprehensive and accurate fuel consumption prediction model.

In future research, in addition to focusing on the aforementioned aspects, researchers should also consider the universality of the chosen models. It is important to conduct a comprehensive fuel consumption analysis by taking into account a variety of vehicle types.
such as passenger cars, trucks, buses, and the performance differences of these vehicle types in different geographic regions (high-altitude areas, mountainous regions, extremely cold regions, etc.).

**Author Contributions:** D.Z.: conceptualization; methodology; visualization. H.L.: writing-original draft; writing-review and editing. J.H.: data curation; writing-review and editing. P.G.: funding acquisition; investigation. Y.Z.: writing-review and editing. W.H.: methodology. Z.F.: methodology. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (Grant No. 62073298), partly supported by the Key Research and Development Projects in Henan Province in 2022 (221111240200), partly supported by the Key Scientific and Technological Project of Henan Province (23210221040), partly supported by the Opening Project of Key Laboratory of operation safety technology on transport vehicles, Ministry of Transport, PRC (KFKT2022-05, and partly funded by the Research Foundation of Zhengzhou University of Light Industry (2021BSJ021).

**Data Availability Statement:** This review is based on the analysis and review of existing research results in the literature, without creating new data sets.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


6. Topić, J.; Škugor, B.; Deur, J. Neural Network-Based Prediction of Vehicle Fuel Consumption Based on Driving Cycle Data. *Sustainability* 2022, 14, 744. [CrossRef]


27. Ziolkowski, J.; Oszczyapa, M.; Malachowski, J.; Szkatnik-Rogoż, J. Use of Artificial Neural Networks to Predict Fuel Consumption on the Basis of Technical Parameters of Vehicles. *Energies* 2021, 14, 2639. [CrossRef]


50. Katreddi, S.; Thiruvengadam, A. Trip Based Modeling of Fuel Consumption in Modern Heavy-Duty Vehicles Using Artificial Intelligence. Energies 2021, 14, 8592. [CrossRef]
53. Panapakidis, I.; Sourti, V.-M.; Dagoumas, A. Forecasting the Fuel Consumption of Passenger Ships with a Combination of Shallow and Deep Learning. Electronics 2020, 9, 776. [CrossRef]
55. Valido, M.R.; Gomez-Cardenes, O.; Magdaleno, E. Monitoring Vehicle Pollution and Fuel Consumption Based on AI Camera System and Gas Emission Estimator Model. Sensors 2023, 23, 312. [CrossRef]
64. Li, Y.; Zhou, S.; Liu, J.; Tong, J.; Dang, J.; Yang, F.; Ouyang, M. Multi-objective optimization of the Atkinson cycle gasoline engine using NSGA III coupled with support vector machine and back-propagation algorithm. Energy 2023, 262, 125262. [CrossRef]
68. Wang, Q.; Zhang, R.; Wang, Y.; Lv, S. Machine Learning-Based Driving Style Identification of Truck Drivers in Open-Pit Mines. Electronics 2020, 9, 19. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.