

Article

Prediction of Truck Fuel Consumption Based on Crossformer-LSTM Characteristic Distillation

Kai Du ¹, Qingqing Shi ², Jingni Song ² , Dan Chen ^{1,*} and Weiyu Liu ^{1,*} ¹ Electronic and Control Engineering, Chang'an University, Xi'an 710064, China; dukai@chd.edu.cn² College of Transportation Engineering, Chang'an University, Xi'an 710064, China; 2022234003@chd.edu.cn (Q.S.); sjn@chd.edu.cn (J.S.)

* Correspondence: danchen@chd.edu.cn (D.C.); liuweiyu@chd.edu.cn (W.L.)

Abstract: With the increasing number of heavy-duty trucks and their high fuel consumption characteristics, reducing fuel costs has become a primary challenge for the freight industry. Consequently, accurately predicting fuel consumption for heavy-duty trucks is crucial. However, existing fuel consumption prediction models still face challenges in terms of prediction accuracy. To address this issue, a model named Cross-LSTM Multi-Feature Distillation (CLMFD) is proposed. The CLMFD model employs the Crossformer model and the LSTM model as teacher and student models, respectively, utilizing multi-layer intermediate features for distillation. Fuel consumption data from a vehicular networking system was used in this study. Initially, the raw data were preprocessed by segmenting it into two-kilometer intervals, calculating sample features, and handling outliers using box plots. Feature selection was then performed using XGBoost. Subsequently, the CLMFD model was applied to predict fuel consumption. Experimental results demonstrate that the CLMFD model significantly outperforms baseline models in prediction performance. Ablation studies further indicate that the CLMFD model effectively integrates the strengths of both the Crossformer and LSTM, exhibiting superior predictive performance. Finally, predictions on data with varying masking rates show that the CLMFD model demonstrates robust performance. These findings validate the reliability and practicality of the CLMFD model, providing strong support for future research in fuel consumption prediction.



Academic Editor: Christos J. Bouras

Received: 13 November 2024

Revised: 20 December 2024

Accepted: 25 December 2024

Published: 31 December 2024

Citation: Du, K.; Shi, Q.; Song, J.; Chen, D.; Liu, W. Prediction of Truck Fuel Consumption Based on Crossformer-LSTM Characteristic Distillation. *Appl. Sci.* **2025**, *15*, 283. <https://doi.org/10.3390/app15010283>

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: fuel consumption forecast; vehicle networking data; feature selection; cross-LSTM multi-feature distillation

1. Introduction

In contemporary society, the rapid expansion of global transportation has exacerbated environmental pollution and greenhouse gas emissions, severely impacting the well-being of residents and the livability of urban areas. Reducing greenhouse gas emissions and mitigating environmental pollution have become prominent topics in academic research [1]. In 2019, global emissions of greenhouse gases reached 59 billion tons, marking a 12% rise from 52.5 billion tons in 2010 [2]. The transportation industry is the second-largest emitter of these gases [3]. Therefore, reducing emissions from this sector has become an urgent priority. In China, road transport, which includes both private and commercial vehicles, was responsible for 86.76% of the carbon emissions in the transportation sector that year. Among these, heavy-duty trucks accounted for 54% of the emissions from road transport [4]. Given this context, improving fuel efficiency in heavy-duty trucks is of great significance. This paper focuses on the development of predictive models for fuel consumption in heavy-duty trucks.

Although significant progress has been made in fuel consumption prediction, particularly with the contributions of neural networks, the widespread use of single models has often led to the underutilization of the potential advantages of combining multiple models, which may limit performance in complex scenarios. Furthermore, existing studies have paid limited attention to lightweight model parameters and improved computational efficiency, both of which are crucial for practical applications. Traditional methods often require a large number of parameters, resulting in complex models that struggle to operate efficiently in resource-constrained environments. Our research employs multi-layer feature distillation technology, which not only enhances predictive performance but also reduces redundant parameters, making the model more practically feasible.

The multi-layer feature distillation model offers significant advantages. Firstly, this model synthesizes the multi-layer feature knowledge from both the teacher and student models, enabling a deep understanding of the input data across various levels of abstraction, thereby capturing complex patterns and relationships more effectively. Secondly, by utilizing a teacher–student architecture, the model achieves parameter efficiency without compromising high performance, reducing computational costs and enhancing operational efficiency. Ultimately, the multi-layer feature distillation model improves prediction accuracy by accurately integrating feature representations from different levels.

Therefore, the CLMFD model is introduced, which integrates the Crossformer and LSTM models as teacher and student models, respectively, leveraging their intermediate features. The Crossformer model captures dependencies between time and different variables, while the LSTM model effectively captures and utilizes long-term dependencies in time series data. The CLMFD model harnesses the strengths of both to improve fuel consumption prediction. When compared with existing fuel consumption prediction models such as BP neural networks, random forests, RNNs, Transformers, and PatchTST, the CLMFD model demonstrates superior performance. The main contributions of this study are as follows:

- The XGBoost feature selection method was employed to extract seven highly relevant features from an initial set of 15, which were then used as inputs for the fuel consumption prediction model.
- The CLMFD model is proposed, which combines the Crossformer and LSTM models as teacher and student, integrating their intermediate features during training. The exceptional performance of the CLMFD model was validated using vehicular network data and through comparison with multiple baseline models, including BP neural networks, random forests, RNNs, Transformers, and PatchTST.
- The Crossformer model accounts for both temporal dependencies and inter-variable relationships during modeling, while the LSTM model effectively captures and utilizes long-term dependencies in sequential data. The combination of these models in CLMFD enhances its predictive capabilities. Furthermore, the CLMFD model demonstrated superior robustness to outlier data compared with baseline models, underscoring its resilience.

The organization of this paper is as follows: Section 2 reviews the relevant literature, Section 3 outlines the proposed methodology, Section 4 describes the data sources and pre-processing techniques, Section 5 presents the results along with a discussion, and Section 6 provides the conclusion.

2. Related Work

Research on fuel consumption prediction can be categorized into studies on the factors influencing fuel consumption and studies on fuel consumption prediction models.

2.1. Study of Factors Affecting Fuel Consumption

In academic research, it is widely acknowledged that fuel consumption and emission levels of vehicles are influenced by various factors in different environments. Ahn et al. (2002) [5] classified these factors into six primary categories: travel-related factors, weather-related factors, vehicle-related factors, road-related factors, traffic-related factors, and driver-related factors. Ben-Chaim et al. [6] demonstrated through controlled experiments that engine power, speed, and fuel type directly impact fuel consumption performance. Zhang et al. [7] investigated the effects of driving speed and road congestion on fuel consumption, finding that fuel consumption is higher under low-speed, congested traffic conditions. Carrese et al. [8] discovered through experiments that rational driving behavior can reduce fuel consumption by up to 27%.

2.2. Research on Fuel Consumption Models

Existing truck fuel consumption prediction methods can be divided into two main types: (1) physical models based on vehicle dynamics principles, and (2) data-driven models [9].

The first type of model primarily employs mathematical formulas based on the internal structure of vehicles and the operating principles of their components, such as the physical or chemical processes within the engine, to provide accurate predictions [10]. For instance, Chang et al. [11] utilized sensors installed along specific road sections to capture vehicle state parameters at designated locations as inputs for their model. Similarly, Huang et al. [12] used traditional microscopic models to predict vehicle fuel consumption. These models are characterized by a highly deterministic mathematical framework, requiring a profound understanding of the system and its critical sub-processes. However, they are often constrained to specific regions and predefined routes, neglecting the impact of varying road conditions and weather, which limits their applicability in real-world scenarios.

The second category of models primarily depends on sensors and onboard equipment to gather comprehensive operational data related to fuel usage. By analyzing extracted features from the data, these models establish nonlinear correlations with fuel consumption, enabling accurate predictions. Data-driven approaches for predicting fuel consumption often incorporate machine learning methods like random forests and SVMs, alongside deep learning techniques such as BP neural networks, RNNs, and LSTMs. For instance, Zeng et al. [13] utilized SVM to build a regression model for predicting fuel consumption, considering factors like driving distance, and achieved results using extensive datasets. Nevertheless, SVM struggles with nonlinear regression problems, leading to reduced accuracy when the connection between variables and fuel consumption lacks linearity. Du et al. [14] applied a BP neural network to develop a model for predicting fuel consumption, which examines temporal and spatial dimensions to better understand the influencing factors. The findings showed that the BP neural network was effective and appropriate for predicting fuel consumption. Additionally, Kanarachos et al. [15] suggested using an RNN for instantaneous fuel consumption predictions; however, conventional RNNs face challenges like gradient explosion and vanishing, which may reduce efficiency. Bougiouklis et al. [16] proposed an LSTM-based energy management strategy for electric vehicles, achieving a 24.03%.

Despite significant advancements in fuel consumption prediction, challenges remain, particularly in addressing missing data and enhancing model practicality. Traditional physics-based models have limitations in accounting for the impact of various road and weather conditions on fuel consumption. Simultaneously, some data-driven models may perform suboptimally when handling nonlinear relationships and complex spatiotemporal patterns.

In this study, distillation techniques have been introduced to improve the practicality of our model, address prediction issues with missing data, optimize processing speed, and reduce parameters. By transferring knowledge from the teacher model to the student model, we enhanced computational efficiency while maintaining predictive performance. This innovative approach provides a viable solution for effective fuel consumption prediction in real-world applications. The CLMFD model not only excels in prediction performance but also improves practical usability, particularly in addressing missing data. This research presents a more innovative and practical method for future fuel consumption prediction tasks.

3. Methodology

3.1. Problem Description

Fuel consumption prediction involves using a multi-dimensional time series (MTS) where each dimension represents a variable and each time step represents a node, with trucks operating over a distance of two kilometers. The MTS can be denoted as $\mathbf{x}_t$:

$$\mathbf{x}_t = [x_{1,t}, x_{2,t}, \dots, x_{D,t}] \quad (1)$$

where $x_{i,t}$ represents the value of the i -th variable related to fuel consumption at time t .

Fuel consumption prediction is a typical time series forecasting problem. The goal is to predict the fuel consumption at the next time step based on the previous n time steps of the multi-dimensional time series $\mathbf{x}_t$, using the CLMFD model represented by F . This prediction process can be expressed by the following formula:

$$\hat{y}_{t+1} = F(\mathbf{x}_{t-n+1}, \mathbf{x}_{t-n+2}, \dots, \mathbf{x}_t) \quad (2)$$

where $\hat{y}_{t+1}$ represents the predicted fuel consumption at the next time step.

3.2. XGBoost

XGBoost [17], an enhanced ensemble learning algorithm, integrates GBDT and RF models, focusing on solving complex nonlinear relationships. Its core idea is to iteratively add trees and split features to progressively build a tree. Each new tree aims to learn a new function to better fit the residuals from the previous round of predictions. The optimization criterion is to minimize the objective function, which is formulated as follows (Equation (3)). The goal of XGBoost is to progressively combine a set of “weak” learners into a “strong” learner, enabling fast and accurate solutions to various data science problems.

$$Obj^{(t)} = \sum_{i=1}^n \left[l(y_i, (\hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))) + \Omega(f_t) \right] + c \quad (3)$$

where n denotes the total number of samples; $\hat{y}_i^{(t-1)}$ represents the estimate at iteration $t - 1$; y_i denotes the observed value of the i -th sample; $f_t(\mathbf{x}_i)$ is the new tree added in iteration t ; l is the loss function between observed values and estimates; $\Omega(f_t)$ represents the regularization term to prevent model overfitting; and c is a constant.

After constructing the boosting trees, the importance of each feature can be conveniently evaluated through gradient boosting. Feature importance scores reflect the contribution of each feature in building the decision trees. A feature’s importance increases with the number of times it is used in the trees. Importance is calculated by assessing the improvement in model performance at each feature’s split point and weighting based on the tree’s node positions. Typically, split points closer to the root node have higher weights. Specific performance metrics can include Gini impurity and other functions. Ultimately,

feature importance scores are derived by averaging the results across all trees, which can be used for ranking and comparison.

3.3. Cross-LSTM Multi-Feature Distillation

After identifying the significant factors influencing fuel consumption with XGBoost, the subsequent step involves forecasting fuel consumption using the Cross-LSTM Multi-Feature Distillation method. In this approach, Crossformer acts as the teacher model, while LSTM operates as the student model. These are linked through a multi-layer feature distillation mechanism. This design not only improves the precision of LSTM predictions but also minimizes parameter usage and enhances runtime efficiency. The detailed structure is depicted in Figure 1.

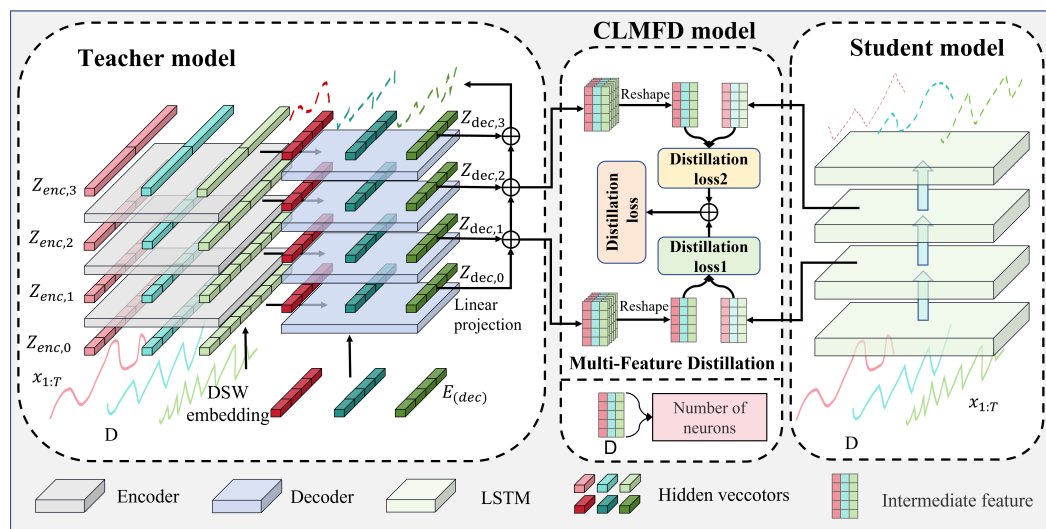


Figure 1. Structure of Cross-LSTM multi-feature distillation.

3.3.1. Crossformer

In this architecture, Crossformer [18] acts as the teacher model and comprises three main components: Dimension-wise Segment-wise Embedding (DSW), a two-level attention mechanism, and a hierarchical encoder–decoder structure. In the encoder, adjacent vectors in the temporal domain are systematically merged and an attention mechanism is applied. In the decoder, each layer takes the encoded array from the previous layer as input and outputs a decoded 2D array. This hierarchical structure enables the model to extract critical features from different scales, thereby enhancing its predictive capability in multivariate time series applications.

3.3.2. LSTM

LSTM [19] is a deep learning model specifically designed to handle and learn from sequence data. Compared with traditional Recurrent Neural Networks (RNNs), LSTM addresses the issues of vanishing and exploding gradients by introducing gating mechanisms, thus better capturing long-term dependencies within sequences. The core idea of LSTM includes three gates—input gate, forget gate, and output gate—along with an internal cell state. These gates control the flow of information, effectively managing long-term dependencies in the sequence. Through its gating mechanisms, LSTM processes long-term dependencies efficiently, enhancing its ability to understand complex structures in time series data.

3.3.3. Feature Distillation

Multi-layer feature distillation [20] is a transfer learning technique that employs Crossformer and LSTM as the teacher and student models, respectively. During the pre-training phase, the Crossformer model transfers its hidden layer feature knowledge to the LSTM model, providing it with rich prior knowledge. In the feature distillation process, both the LSTM student model and the Crossformer teacher model receive complete data inputs, which helps ensure that the teacher model provides high-quality guidance. The final error consists of two key components: the student model prediction error and the feature distillation error. This combination of teacher and student models through feature distillation allows the student model to maintain high performance while achieving a more lightweight structure and faster inference speed.

The student model prediction error, which reflects the accuracy of the model's predictions, is given by the following formula:

$$\text{Pre Loss} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (4)$$

where N represents the total number of samples; $\hat{y}_i$ denotes the predicted value for the i -th sample; and y_i represents the true value for the i -th sample.

The feature distillation error (FD Loss) measures the consistency of feature representations between the teacher and student models and is expressed as

$$\text{FD Loss} = \frac{1}{M} \sum_{i=1}^D \|\phi_i(T) - \psi_i(S)\|_2^2 \quad (5)$$

where M denotes the number of hidden layers used in the distillation; $\phi_i(T)$ represents the output of the i -th hidden layer of the teacher model (Crossformer); and $\psi_i(S)$ represents the output of the i -th hidden layer of the student model (LSTM). A smaller distillation error indicates that the student model has better learned the knowledge from the teacher model.

Thus, the total loss for the student model during prediction is

$$T\text{Loss} = \text{Pre Loss} + \text{FD Loss} \quad (6)$$

4. Data Source and Preprocessing

4.1. Data Source

The data used in this study come from the heavy truck data of the car networking system of a heavy automobile company. Relevant information about the collected data is shown in Table 1. The data from the Xi'an–Yangshan route are used as training data, while the data from the Xi'an–Hanzhong and Xi'an–Baomao routes are used as the test set. A total of 534,687 records were collected with a sampling interval of 5 s. Each record includes parameters such as Vehicle ID, Timestamp, Longitude, Latitude, Altitude, Speed (km/h), Engine RPM (rpm), Mileage (km), and Cumulative Fuel Consumption, totaling 81 parameters. Sample data are shown in Table 2.

Table 1. Collected Data.

Vehicle ID	Route	Date
Vehicle A	Xi'an–Shanyang	1 January 2021–30 January 2021
Vehicle B	Xi'an–Shanyang	26 December 2020, 29 December 2020
Vehicle C	Xi'an–Hanzhong	1 January 2021, 15 January 2021, 19 January 2021
Vehicle D	Xi'an–Baomao	2 January 2021, 5 January 2021

Table 2. Vehicle data sample.

Vehicle ID	Timestamp	Altitude (m)	Speed (km/h)	RPM	Cumulative Fuel (L)	Mileage (km)	...
Vehicle A	6 January 2021 9:12:32	297	33	1160	100,742	238,172.3	...
Vehicle A	6 January 2021 9:12:37	295	35	1160	100,743	238,172.3	...
Vehicle A	6 January 2021 9:12:42	301	30	870	100,743	238,172.3	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Vehicle A	6 January 2021 9:22:47	317	39	1150	100,743	238,176.2	...
Vehicle A	6 January 2021 9:22:52	319	41	1130	100,744	238,176.2	...

4.2. Data Preprocessing

In the raw data, the mileage (km) represents the total mileage accumulated by the vehicle from the start to the present, with a resolution of 0.1 km; the cumulative fuel consumption represents the total fuel consumed by the vehicle from the start to the present, with a resolution of 1 L. Due to the relatively large resolution of the collected data, situations may occur where the cumulative fuel consumption and mileage are the same at different times. To calculate the fuel usage per record, it is necessary to refine the cumulative fuel data. The approach used in this study involves determining the vehicle data corresponding to a one-liter increase in cumulative fuel. Assuming a uniform distribution of fuel consumption, the amount of fuel used by the vehicle every 5 s can then be computed.

4.2.1. Data Partitioning

Samples are obtained by partitioning the mileage attribute in the collected data, with each sample corresponding to every two kilometers. Features for each sample are calculated, and a summary of these features is provided in Table 3. After partitioning the collected data, there are a total of 14,472 samples, each containing 16 features. The partitioned samples are sequential data and can be analyzed using time series methods.

Table 3. Feature Description.

Feature (Abbreviation)	Description
Braking Count (BC)	Total number of braking events within two kilometers
RPM	Average RPM within two kilometers
Average Altitude (AA)	Average altitude within two kilometers
Uphill (Uh)	Average uphill height within two kilometers
Downhill (Dh)	Average downhill height within two kilometers
Peak Speed (PS)	Maximum speed within two kilometers
Bottom Speed (BS)	Minimum speed within two kilometers
Average Speed (AS)	Average speed within two kilometers
Speed Standard Deviation (SSD)	Standard deviation of speed within two kilometers
Acceleration Time (AT)	Acceleration time within two kilometers
Deceleration Time (DT)	Deceleration time within two kilometers
Cruise Time (CT)	Cruise time within two kilometers
Acceleration Share (AS)	Share of acceleration time relative to total time within two kilometers
Deceleration Share (DS)	Share of deceleration time relative to total time within two kilometers
Cruise Share (CS)	Share of cruise time relative to total time within two kilometers
Fuel Consumption (FC)	Fuel consumption within two kilometers

4.2.2. Outlier Handling

The following data processing methods were employed to ensure the accuracy of model training and prediction: First, we removed missing values from the dataset to maintain data integrity and consistency. Following this step, the dataset contains 14,439 samples. Second, for excessively large or small outliers, we used box plots to analyze each feature. The box plot calculates the quartiles (Q1, Q3) and interquartile range (IQR) to determine

the upper and lower bounds for outliers. Specifically, the outlier bounds are calculated as follows:

$$\text{Upper Bound} = Q3 + 1.5 \times IQR \quad (7)$$

$$\text{Lower Bound} = Q1 - 1.5 \times IQR \quad (8)$$

In the data preprocessing step, values falling outside of the established bounds were adjusted to the corresponding upper or lower limits. This was performed to restrict outliers within a reasonable range, minimizing their influence on model training and prediction. It is important to note that, for outlier detection, we define outliers as extreme values that significantly deviate from the normal data distribution. For instance, an excessively high number of braking events is often associated with equipment malfunction or other abnormal conditions. However, occasional braking due to traffic accidents or congestion is not considered an outlier, as it is temporary and does not affect the general pattern of braking frequency observed in the two-kilometer intervals. Therefore, such cases are not excluded and do not negatively impact the model.

To further visually demonstrate the outliers in the data, we plotted box plots to show the distribution of each feature. Given the large volume of data and the number of features, displaying box plots for all features may result in overly complex visuals. Therefore, we selected four key features—PS, BS, MS, and SSD—and presented box plots for 30 data points of each feature, as shown in Figure 2. The first subplot displays the box plot for the raw data, while the second subplot shows the box plot after outlier handling. These plots effectively illustrate the distribution of the data and the impact of outlier treatment, helping to better understand the data structure and the outlier correction process during model training.

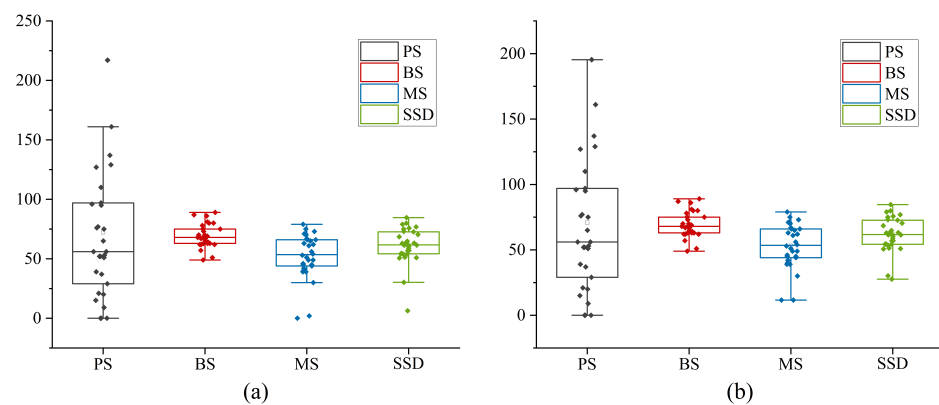


Figure 2. Box diagram of some feature samples. (a) represents the box plot before processing; (b) represents the box plot after processing.

4.2.3. Feature Selection

The purpose of feature selection is to improve model performance and efficiency. The original dataset might contain unnecessary or unrelated features that do not contribute to model training and prediction but increase computational burden and model complexity. By selecting features that significantly impact the target variable from the original feature set, the number of features can be reduced, simplifying the model and improving training and prediction speed.

The feature importance analysis results from the XGBoost model are shown in Figure 3. According to this analysis, cruise time is the most important feature with a contribution rate of 29.25%, followed by braking count. Considering the relationship between the number of features and the model Mean Squared Error (MSE), as shown in Figure 3, it is observed that

the MSE remains unchanged when the number of features exceeds 7. Therefore, we select the seven features with the highest contribution for the subsequent prediction model input: cruise time, braking count, average speed, downhill, uphill, RPM, and average altitude.

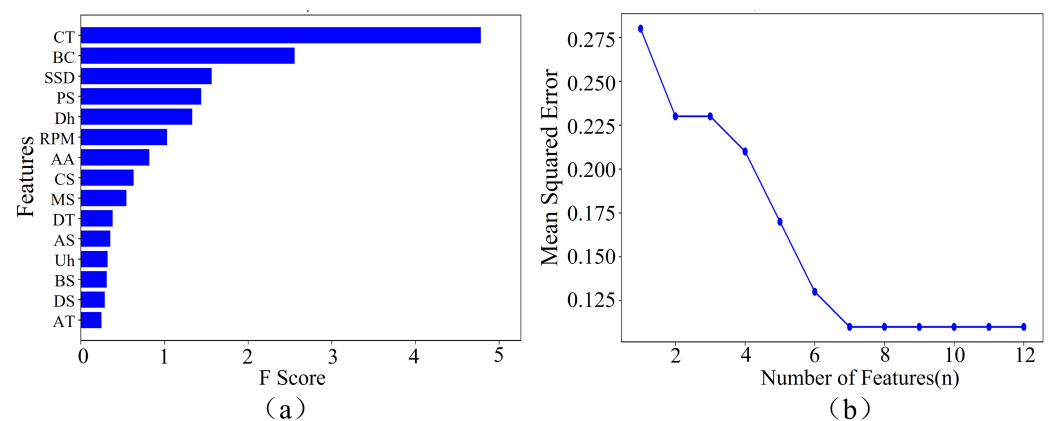


Figure 3. The results of XGBoost. (a) Feature Importance plot; (b) Mean Squared Error vs. Number of Features plot.

In addition, to validate the robustness of the selected features, we conducted a feature selection experiment based on SHAP values. The results are shown in Figure 4. The left chart presents a bar plot based on feature importance analysis, with features ordered by their importance scores from high to low. The feature names are listed on the vertical axis, and the horizontal axis represents the importance scores, providing a clear visualization of the contribution of each feature to the target variable. The right chart displays the feature contribution distribution based on SHAP values, illustrating the positive and negative impacts of each feature on the prediction results. Each row corresponds to a feature, with the horizontal axis representing the magnitude of the SHAP values. The color differentiates between positive and negative contributions (red for positive, blue for negative), and the density and position of the points reflect the relationship between the feature values and the model output.

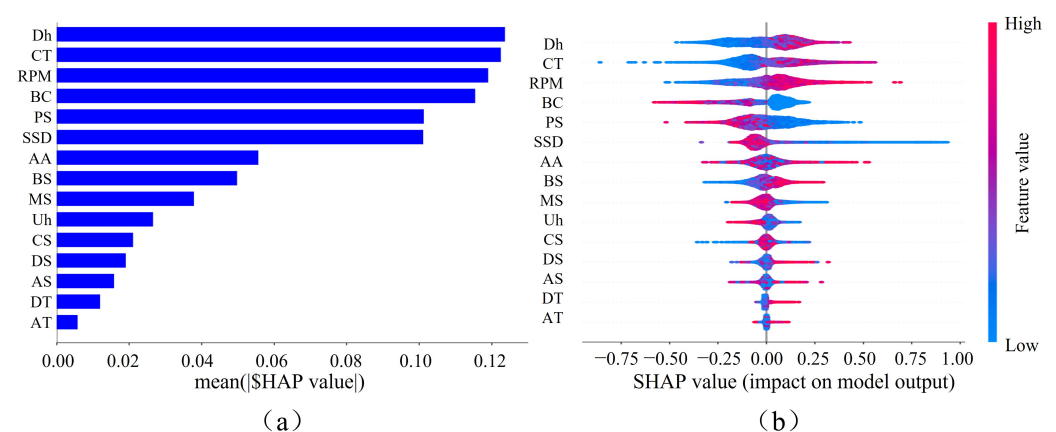


Figure 4. Visualization diagram of feature importance and contribution distribution based on SHAP. (a) represents the feature importance obtained using the SHAP method; (b) visualizes the impact of each sample on fuel consumption.

Compared with the feature importance analysis based on XGBoost, although there are some differences in the ranking of features, the top seven features are consistent. These include cruising time, braking count, average speed, downhill, uphill, rpm, and average

altitude. The heatmap results shown in Figure 5 further demonstrate that the selected features are strongly correlated with fuel consumption, confirming the importance of these features. This suggests that the XGBoost feature selection method is reliable, and the selected features exhibit strong robustness and consistency, making them important variables for input into subsequent models.

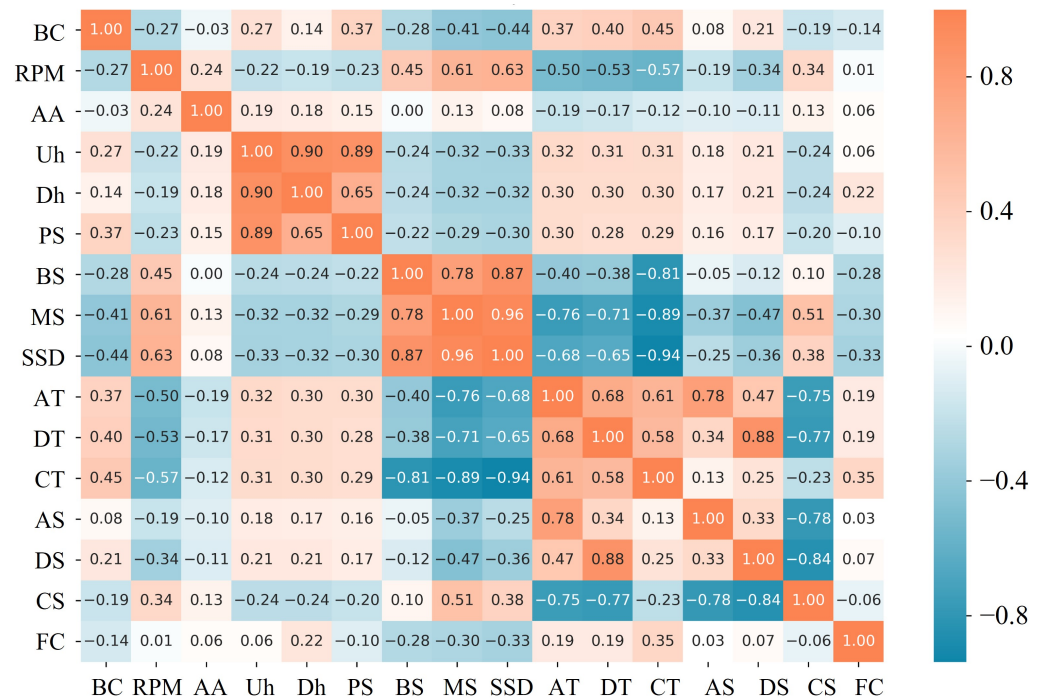


Figure 5. Feature correlation heatmap.

5. Results and Discussion

5.1. Experimental Parameters and Evaluation Metrics

In this study, the CLMFD model was trained using data from the Xi'an to Shanyang route and tested on data from the Xi'an to Hanzhong and Xi'an to Baomao routes. To ensure a fair comparison across different models, several parameters were standardized, including the Adam optimizer, a learning rate of 0.001, and the use of Mean Squared Error (MSE) as the loss function, with training conducted over 100 epochs and a batch size of 32. Additionally, other parameters were adjusted based on the specific characteristics of each model. The detailed parameter settings for each model are as follows: the BP neural network uses a 7×128 , 128×1 architecture; the random forest (RF) model is configured with 100 trees, a maximum depth of 10, square root of the total number of features for max features, minimum samples for splitting set to 2, and minimum samples for leaf nodes set to 1; the RNN uses a 7×128 , 128×1 architecture; the Transformer is configured with 2 encoder layers, 4 attention heads, and a hidden layer size of 128; PatchTST uses a time window size of 48, a patch length of 16, 3 encoder layers, a hidden layer size of 128, and 4 attention heads; in the CLMFD model, both the decoder layer of the teacher model and the network structure of the student model consist of 4 layers.

To comprehensively evaluate the performance of the fuel consumption prediction model, this study employed four commonly used evaluation metrics: MSE, RMSE, MAE, and R^2 . In the experiments, the Python programming language was used, along with popular libraries such as pandas, numpy, XGBoost, scikit-learn, PyTorch, and TensorFlow for data processing, feature selection, and model construction.

5.2. Selection of Multi-Layer Feature Distillation

Since the LSTM model learns knowledge from the Crossformer model through multi-layer feature distillation, selecting which decoder layer outputs to use as intermediate features for learning becomes crucial. To address this issue, this study trains the model using combinations of intermediate features from different layers to identify the optimal intermediate feature distillation combination.

The Crossformer model consists of four hidden layers, with the output of the fourth hidden layer serving as the final prediction result. Therefore, when combining intermediate features, this study only considers the outputs of the first three hidden layers, resulting in seven possible combinations (for example, CLMFD(23) indicates using the outputs of the second and third hidden layers as intermediate features).

To ensure the reliability and stability of the experimental results, all experiments in this study were conducted with multiple independent trials (each group of experiments repeated five times), and the final results were averaged. As shown in Table 4, among the seven combinations, CLMFD(23) performs the best, indicating that the outputs of the second and third hidden layers contain more representative feature information. Therefore, this study selects the outputs of the second and third hidden layers as intermediate features for distillation with the student model.

Table 4. Results of Different Hidden Layers as Intermediate Features.

Feature Combination	MSE	RMSE	MAE	R ²
CLMFD(1)	0.1055	0.3248	0.2102	0.6988
CLMFD(2)	0.0977	0.3125	0.2068	0.7211
CLMFD(3)	0.1043	0.3231	0.2134	0.7020
CLMFD(12)	0.1048	0.3237	0.2212	0.7008
CLMFD(13)	0.1027	0.3205	0.2071	0.7067
CLMFD(23)	0.0962	0.3102	0.2078	0.6716
CLMFD(123)	0.1053	0.3245	0.2129	0.6993

5.3. Model Comparison

The data in Table 5 show significant differences among the models in terms of parameter count and prediction time. The RF model has the shortest prediction time, while the BP neural network model has the fewest parameters. In contrast, the CLMFD model has the highest number of parameters, but its prediction time is second only to the RF model, outperforming the other models.

Table 5. The running time and number of parameters for each model.

Model	Number of Parameters (Million)	Prediction Time (Seconds)
BP neural network	0.02	0.18
RF	1.40	0.06
RNN	0.07	0.32
Transformer	0.15	4.06
PatchTST	0.41	7.18
CLMFD	13.05	0.13

The results in Table 6 demonstrate that the CLMFD model excels across various evaluation metrics (e.g., MAE, RMSE), showcasing superior predictive performance and highlighting its exceptional modeling capability and generalization ability. Moreover, to ensure the reliability and stability of the experimental results, all experiments were

conducted multiple times independently, and the average values were taken to minimize the impact of random factors on performance evaluation.

Table 6. Model Evaluation Results.

Model	MSE	RMSE	MAE	R ²
BP Neural Network	0.1661	0.4076	0.2935	0.5256
RF	0.1347	0.3671	0.2700	0.6152
RNN	0.1210	0.3478	0.2418	0.6546
Transformer	0.1071	0.3288	0.2272	0.6942
PatchTST	0.1442	0.3797	0.2788	0.5883
CLMFD	0.1020	0.3194	0.2101	0.7087

The two subgraphs in Figure 6, respectively, show the absolute errors between the predicted values and the true values of each prediction model in 40 samples. In order to clearly compare the performance of different models, the first sub-graph shows the first 20 samples, and the second sub-graph shows the last 20 samples. As a whole, it can be seen that the error of the CLMFD model is the smallest, and the predicted value is the closest to the real value.

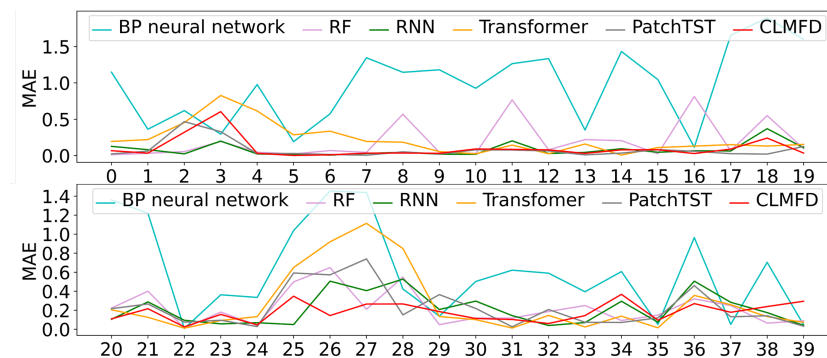


Figure 6. Absolute error between predicted value and true value of each prediction model.

Figures 7 and 8, respectively, present the comparison between the predicted and actual fuel consumption values of the CLMFD model on the Xi'an–Shanyang and Xi'an–Baomao routes. From the figures, it can be observed that the CLMFD model maintains a high level of prediction accuracy on both routes.

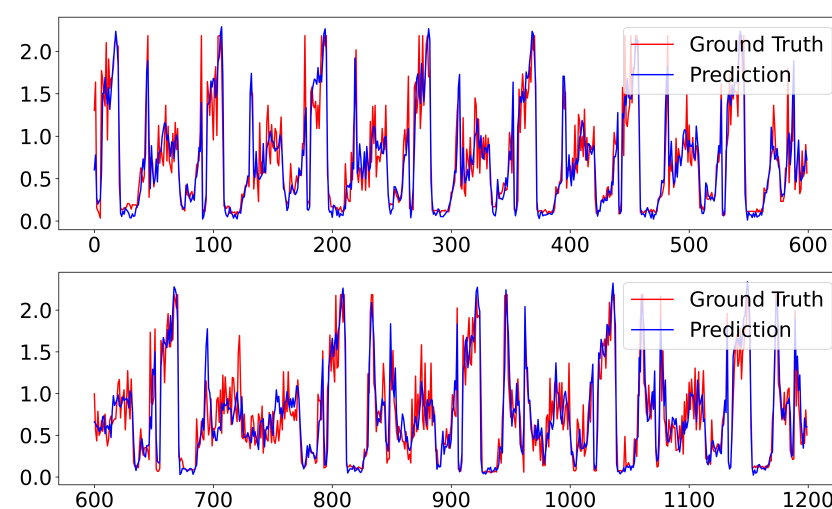


Figure 7. Evaluation results of Xi'an–Shanyang.

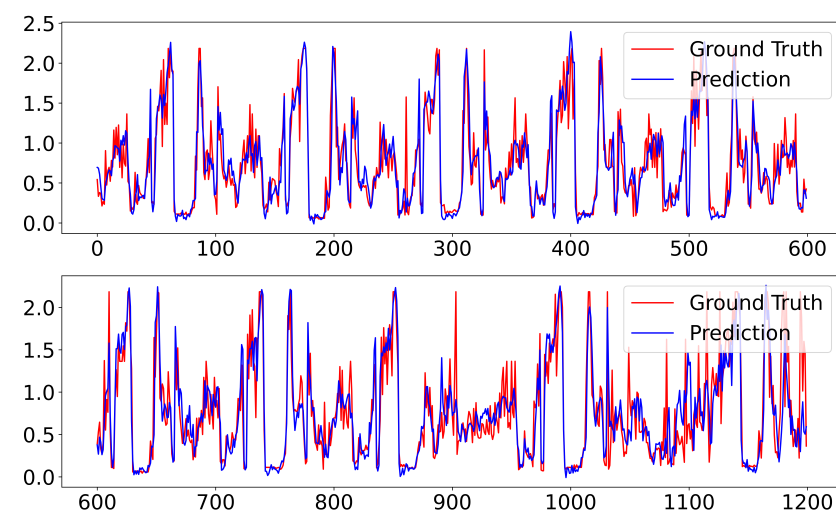


Figure 8. Evaluation results of Xi'an–Baomao.

5.4. Ablation Study

The CLMFD model primarily consists of two components: the teacher model and the student model. The objective of the ablation study is to evaluate the contribution of each module to the overall performance of the model and analyze the effectiveness of the teacher and student models across different metrics. This aims to validate the rationality of the CLMFD model design and identify the sources of its performance improvements. The results of the ablation study are presented in Table 7.

Table 7. Ablation Study Results for the CLMFD Model.

Prediction Method	MSE	RMSE	MAE	R ²
Crossformer	0.0932	0.3159	0.2107	0.7150
LSTM	0.1005	0.3172	0.2070	0.7128
CLMFD	0.0962	0.3102	0.2078	0.6716

The CLMFD model significantly outperforms both the Crossformer model and the LSTM model in terms of MSE and RMSE, reflecting its superior performance in accuracy, particularly in capturing the volatility and trends in fuel consumption prediction. Additionally, the MAE of the CLMFD model is similar to that of the LSTM model and is notably better than the Crossformer model, highlighting the CLMFD model's robustness in terms of average absolute error. Although the R-squared value for the CLMFD model is slightly lower than that of the Crossformer and LSTM models, the CLMFD model remains competitive when considering robustness and accuracy. This indicates that the CLMFD model demonstrates excellent performance in fuel consumption prediction and that the LSTM model effectively learns from the knowledge of the Crossformer model.

5.5. Model Robustness Validation

To verify the robustness of the CLMFD model in the presence of various uncertainties, noise, and anomalous data, random masking was applied to the data (Xi'an to Shanyang; Xi'an to Bao Mao) to simulate anomalous conditions. Anomalous data samples refer to values in the dataset that are introduced due to factors such as data collection errors, transmission interference, noise, and sudden events (e.g., traffic accidents or extreme braking behaviors). These types of data may interfere with the model's training and prediction, affecting its practical application. By randomly masking a portion of the input data samples (i.e., setting different masking rates), we systematically simulate missing data

or anomalous values, assessing the model's robustness and stability when data quality deteriorates, noise increases, and anomalous samples are present. The random masking rates were set to 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40. By gradually increasing the masking rate, changes in model performance were observed using the MSE, MAE, R^2 , and RMSE metrics to verify the CLMFD model's ability to resist interference.

The experimental results, as shown in Figure 9, indicate that the CLMFD model outperforms the baseline models in terms of MSE, MAE, and R^2 . In terms of RMSE, the performance of the CLMFD model is comparable to that of the Transformer model but better than other baseline models. As the masking rate increases, the changes in the evaluation metrics for CLMFD are minimal, indicating that the model exhibits considerable stability. Among the baseline models, the RF model performs the worst. When the masking rate exceeds 0.25, the R^2 value becomes negative, indicating that the model's fit is worse than simply using the average of the target variable, meaning the model's predictions are worse than random guessing. This suggests that the RF model may suffer from significant underfitting.

From the above experiments, it is evident that the CLMFD model exhibits strong robustness and excellent predictive capabilities even in the presence of anomalous data. To further investigate the predictive capability of different components of the CLMFD model under anomalous data, tests were conducted with a masking rate of 0.05. The results, presented in Table 8, show that the CLMFD model outperforms both the Crossformer and LSTM models.

Table 8. Prediction Results of CLMFD Model and Its Components on Anomalous Data.

Model	MSE	RMSE	MAE	R^2
Crossformer	0.1051	0.3243	0.2105	0.6981
LSTM	0.1227	0.3502	0.2245	0.6498
CLMFD	0.1019	0.3065	0.2065	0.7319

By integrating the advantages of both models through multi-layer feature distillation, the CLMFD model improves generalization on unseen data and helps the student model learn more information from the teacher model, addressing anomalous data issues and enhancing model prediction performance.

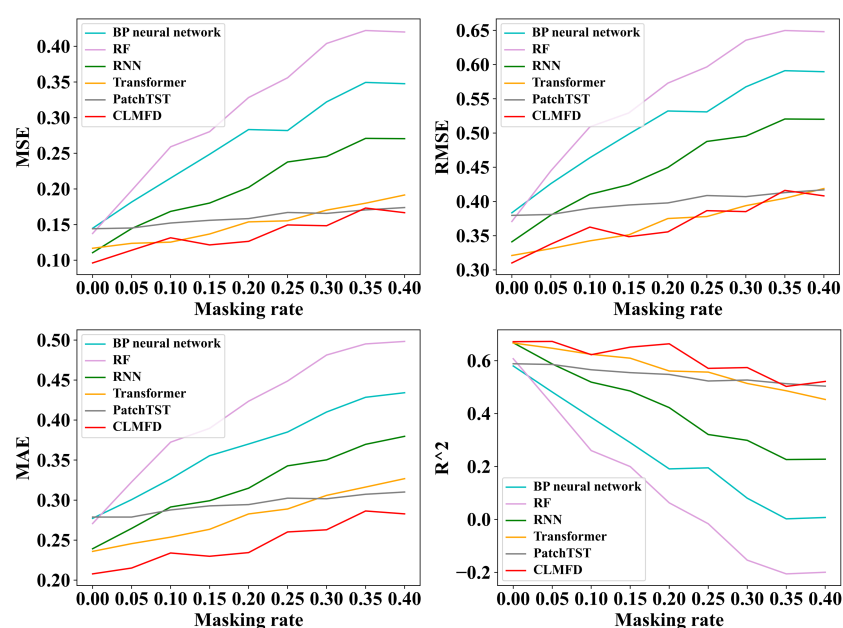


Figure 9. Prediction results of abnormal data with different mask rates.

6. Conclusions

This study is based on truck driving data from a vehicle networking system of a certain group. The raw data were preprocessed by dividing the driving data into samples every two kilometers and calculating various features for each sample. Using the XGBoost feature selection method, seven key features—cruise time, braking count, average speed, downhill, uphill, RPM, and average altitude—were identified as the main factors influencing fuel consumption. Based on these, a fuel consumption prediction model was developed using CLMFD and compared with BP neural networks, RF, RNN, Transformer, and PatchTST models.

Experimental results show that the CLMFD model achieves the best predictive performance, with an MSE of 0.0962, significantly outperforming the other models. In particular, when the data contains outliers, the CLMFD model effectively transfers knowledge from the Crossformer model to LSTM, enabling LSTM to better adapt to complex time series data, thereby improving the model's accuracy and robustness.

The proposed CLMFD model not only excels in complex time series prediction tasks but also enhances the model's generalization capability through multi-feature distillation. It is suitable for multi-scenario, multi-variable time series tasks and demonstrates strong scalability. However, there are some limitations in practical applications: firstly, the model's complexity, parameter count, and computational time may pose challenges for real-world deployment, and secondly, the model is sensitive to hyperparameter selection, which may require additional tuning to achieve optimal performance; moreover, the current study has not been extensively validated in low-computation devices or real-time scenarios, and further optimization is needed for practical use.

In future research, we will expand the dataset and incorporate additional features, including road conditions, traffic patterns, and weather factors, to better understand their impact on fuel consumption. Moreover, since the current dataset includes only a limited number of vehicle types, the variations in fuel consumption across different models have yet to be explored. Future work will also focus on developing fuel consumption prediction models tailored to specific vehicle types, thereby improving the model's versatility and providing more precise support for vehicle energy efficiency and emission reduction efforts.

Author Contributions: Conceptualization, K.D. and J.S.; formal analysis, D.C. and W.L.; methodology, Q.S.; resources, K.D. and W.L.; supervision, J.S.; validation, D.C.; writing—original draft, Q.S.; writing—review and editing, K.D., J.S., D.C. and W.L. 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 under grant 52202385, and the Anhui Provincial Key Laboratory of Urban Rail Transit Safety and Emergency Management, Hefei University under grant 2024GD0009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: This research benefited greatly from the contributions of Letian Li and Yanming He, whose valuable revisions and feedback during the revision stage significantly improved the manuscript.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Sumaira; Siddique, H.M.A. Industrialization, energy consumption, and environmental pollution: Evidence from South Asia. *Environ. Sci. Pollut. Res.* **2023**, *30*, 4094–4102. [[CrossRef](#)]
2. Shukla, P.R. *Climate Change 2022: Mitigation of Climate Change: Working Group III Contribution to the 6th Assessment Report of the Intergovernmental Panel on Climate Change: Summary for Policymakers*; IPCC: Geneva, Switzerland, 2022.
3. Jafar, H.A.; Shahrouh, I.; Mroueh, H. Use of a Hybrid Approach to Estimate Greenhouse Gas Emissions from the Transport Sector in Palestine. *Climate* **2023**, *11*, 170. [[CrossRef](#)]
4. Li, X.; Tan, X.; Wu, R.; Xu, H. Research on Carbon Peaking and Carbon Neutrality Paths in the Transportation Field. *China Eng. Sci.* **2021**, *23*, 15–21.
5. Trani, A.; Rakha, H.; Ahn, K. Estimating Vehicle Fuel Consumption and Emissions based on Instantaneous Speed and Acceleration Levels. *J. Transp. Eng.* **2002**, *18*, 182–190.
6. Ben-Chaim, M.; Shmerling, E.; Kuperman, A. Analytic modeling of vehicle fuel consumption. *Energies* **2013**, *6*, 117–127. [[CrossRef](#)]
7. Zhang, S.; Wu, Y.; Liu, H.; Huang, R.; Un, P.; Zhou, Y.; Fu, L.; Hao, J. Real-world fuel consumption and CO₂ (carbon dioxide) emissions by driving conditions for light-duty passenger vehicles in China. *Energy* **2014**, *69*, 247–257. [[CrossRef](#)]
8. Carrese, S.; Gemma, A.; La Spada, S. Impacts of driving behaviors, slope, and vehicle load factor on bus fuel consumption and emissions: A real case study in the city of Rome. *Procedia-Soc. Behav. Sci.* **2013**, *87*, 211–221. [[CrossRef](#)]
9. Zhao, D.; Li, H.; Hou, J.; Gong, P.; Zhong, Y.; He, W.; Fu, Z. A Review of the Data-Driven Prediction Method of Vehicle Fuel Consumption. *Energies* **2023**, *16*, 5258. [[CrossRef](#)]
10. Wang, J.; Rakha, H.A. Fuel Consumption Model for Heavy Duty Diesel Trucks: Model Development and Testing. *Transp. Res. Part Transp. Environ.* **2017**, *55*, 127–141. [[CrossRef](#)]
11. Chang, X.; Chen, B.Y.; Li, Q.; Cui, X.; Tang, L.; Liu, C. Estimating Real-time Traffic Carbon Dioxide Emissions Based on Intelligent Transportation System Technologies. *IEEE Trans. Intell. Transp. Syst.* **2012**, *14*, 469–479. [[CrossRef](#)]
12. Huang, W.; Guo, Y.; Xu, X. Evaluation of Real-time Vehicle Energy Consumption and Related Emissions in China: A Case Study of the Guangdong–Hong Kong–Macao Greater Bay Area. *J. Clean. Prod.* **2020**, *263*, 121583. [[CrossRef](#)]
13. Zeng, W.; Miwa, T.; Morikawa, T. Exploring trip fuel consumption by machine learning from GPS and CAN bus data. *J. East. Asia Soc. Transp. Stud.* **2015**, *11*, 906–921.
14. Du, Y.; Wu, J.; Yang, S.; Zhou, L. Predicting vehicle fuel consumption patterns using floating vehicle data. *J. Environ. Sci.* **2017**, *59*, 24–29. [[CrossRef](#)] [[PubMed](#)]
15. Kanarachos, S.; Mathew, J.; Fitzpatrick, M.E. Instantaneous vehicle fuel consumption estimation using smartphones and recurrent neural networks. *Expert Syst. Appl.* **2019**, *120*, 436–447. [[CrossRef](#)]
16. Bougiouklis, A.; Korkofigkas, A.; Stamou, G. Improving fuel economy with LSTM networks and reinforcement learning. In Proceedings of the 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; pp. 230–239.
17. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
18. Zhang, Y.; Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In Proceedings of the 11th International Conference on Learning Representations, Online, 5–29 April 2023.
19. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
20. Gou, J.; Xiong, X.; Yu, B.; Du, L.; Zhan, Y.; Tao, D. Multi-target knowledge distillation via student self-reflection. *Int. J. Comput. Vis.* **2023**, *131*, 1857–1874. [[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.