ISSN: 1813-4890

Research on an Automotive Data Integration System Based on Machine Learning

Tian Gao[1], Xuehong Wang [2], Jianming Li [2,*]

[1] CATARC Automotive Inspection Center (Wuhan) Co.,Ltd, Wuhan, China

[2]China Auto Information Technology (Tianjin) Co., Ltd. Tianjin, China

Corresponding author: lijianming@catarc.ac.cn

Abstract

With the transformation of the automotive industry toward intelligence and connectivity, multi-source data generated from onboard sensors, 4S dealership management systems, user behavior platforms, and transportation networks have grown explosively. These data are characterized by dispersed sources, heterogeneous types (structured, semi-structured, and unstructured), and inconsistent quality, which makes it difficult to effectively exploit their potential value. To address these challenges, this paper proposes an automotive data integration system based on machine learning. The system consists of four core modules—data acquisition, data cleaning, data fusion, and data mining—and incorporates machine learning algorithms such as anomaly detection and missing value detection to achieve standardized and consistent integration of multi-source automotive data. Experimental results demonstrate that the proposed system provides high-quality data support for intelligent driving, vehicle condition prediction, and user behavior analysis.

Keywords

Machine Learning; Automotive Data; Data Integration; Multi-source Heterogeneous Data; Anomaly Detection; Data Fusion.

1. Introduction

In recent years, with the widespread adoption of intelligent connected vehicles and new energy vehicles, the scale of automotive data has grown exponentially. A single intelligent vehicle can generate terabytes of data per hour, including onboard sensor data, environmental perception data, user behavior data, and server-side operational data. By 2025, the automotive industry has become the third most data-intensive sector worldwide, following finance and healthcare.

The demand for data-driven value creation in the automotive industry is increasingly urgent. On the manufacturing side, data can be leveraged to optimize production processes and reduce failure rates. On the service side, personalized maintenance and roadside assistance services can be provided through data analytics. For intelligent driving, high-quality integrated datasets are essential for training perception models and improving decision-making accuracy. On the operational side, user behavior data can be utilized to refine product design and expand value-added services.

Although automotive data will play a decisive role in future vehicle research and development, several critical challenges remain. First, data sources are highly fragmented and lack a unified access platform, with information distributed across onboard terminals, enterprise servers, and third-party systems. Second, the data exhibit strong heterogeneity, including structured, semi-structured, and unstructured formats. Third, data quality is often poor due to missing values, anomalies, and inconsistencies.

ISSN: 1813-4890

To address these issues, integration of multi-source heterogeneous data is required. Multi-source data fusion technologies have achieved promising results across various domains. For example, Ref. [1] proposes a detection scheme based on multi-source data fusion to address false positives and missed detections in charging pile anomaly identification. The method integrates operational measurements, control data, power grid information, and safety monitoring data to extract behavioral features, which are then matched with user profiles for anomaly scoring. Experimental results show that both false positive and false negative rates are controlled within 1%, significantly improving detection accuracy.

Ref. [2] presents an intelligent transportation system integrating multi-source data fusion and decision support technologies to mitigate urban traffic congestion and high accident rates. The system adopts a hierarchical fusion architecture, where an improved Kalman filtering algorithm processes sensor data at the lower layer, deep neural networks extract key features at the intermediate layer, and Bayesian inference performs decision fusion at the upper layer. The system achieves high-precision traffic flow identification, millisecond-level real-time decision-making, and significantly improved traffic efficiency in practical deployments.

Ref. [3] emphasizes the importance of constructing an intelligence perception system based on multi-source data fusion to address uncertainties in industrial competition and complex information environments. Guided by intelligence perception theory, the proposed framework integrates data perception, contextual awareness, situational awareness, and intelligence response, providing strategic support for accurate, dynamic, and intelligent decision-making in emerging industries.

In summary, this study proposes a machine learning-based automotive data integration system that cleans low-quality data, connects multi-dimensional automotive datasets, and enables enterprises to extract high-value insights from integrated data.

2. Research Framework

The research framework of this study consists of three components: data cleaning, data fusion, and data mining.

Data cleaning aims to address low data quality issues, primarily by detecting and handling anomalies and missing values. Data fusion integrates structured, semi-structured, and unstructured data into a unified representation. Data mining leverages the fused data in combination with machine learning models to extract high-quality and actionable information.

The overall research framework is illustrated in Fig. 1.

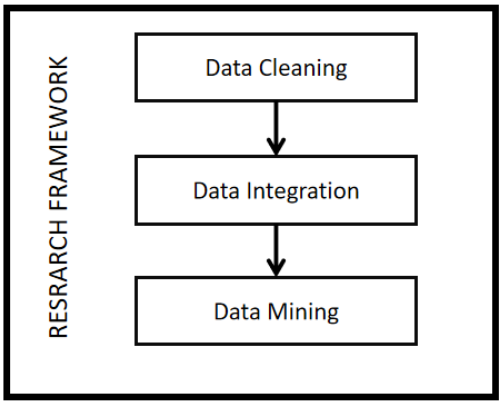

Fig 1 Research Framework

ISSN: 1813-4890

2.1. Data Cleaning

The data cleaning process is divided into several stages, including data source identification and classification, data deduplication, anomaly definition, anomaly processing, missing value localization, and missing value treatment [4–5].

The overall structure of the data cleaning procedure is illustrated in Fig. 2.

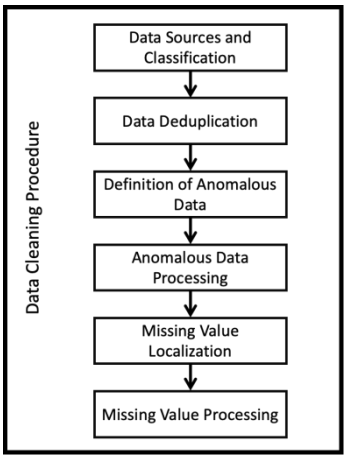

Fig 2 Data Cleaning Procedure

2.1.1. Data Sources and Classification

Automotive data originate from diverse sources and can generally be categorized into onboard terminal data, user behavior data, and manufacturing data.

Onboard terminal data are generated by in-vehicle sensors, OBD interfaces, and infotainment systems, including variables such as acceleration, angular velocity, tire pressure, engine speed, and fuel consumption.

User behavior data are collected from automotive applications, navigation systems, and onboard entertainment platforms, covering information such as driving routes, charging habits, and music preferences.

Manufacturing data are obtained from enterprise production systems and include component parameters, assembly processes, and quality inspection records.

2.1.2. Data Deduplication

Data deduplication refers to the process of identifying and removing or merging duplicate records during data processing to ensure data uniqueness and accuracy.

Common deduplication approaches include key-field-based methods, similarity-based methods, hash-based methods, and database-based methods. In this study, a hash-based deduplication method is adopted.

The fundamental principle of hash-based deduplication is to convert the content of each record into a unique hash value and determine duplication efficiently by comparing hash values.

The mathematical logic of the hash-based deduplication process is as follows:

(1) Compute the hash value h(x) for each data record x;

(2) Store the hash value in a hash table (e.g., a dictionary or an array);

(3) When processing new data, compute its hash value. If the value already exists in the hash table, the record is identified as duplicate; otherwise, it is inserted into the hash table.

2.1.3. Definition of Anomalous Data

Anomaly handling is a critical step in data preprocessing. It aims to identify and correct extreme values that deviate significantly from the normal range to ensure the reliability of subsequent analysis.

ISSN: 1813-4890

Anomalies are typically detected based on the statistical distribution characteristics of the dataset [6–7]. Common methods include the interquartile range (IQR) method and the Z-score method. In this study, the Z-score method is adopted.

The procedure of the Z-score method is described as follows:

(1) Z-score calculation

For a data point *X*, the Z-score is calculated as shown in Formula (1).

$$X_Z = \frac{X - X_\mu}{X_\sigma} \tag{1}$$

where *Z* denotes the standardized value, μ represents the mean of the dataset, and σ denotes the standard deviation.

(2) Anomaly identification

A threshold is first defined. Data points with an absolute Z-score greater than 3 (i.e., $|Z|>3$) are regarded as anomalous values.

2.1.4. Anomalous Data Processing

Anomaly processing can be categorized into two approaches: direct handling and statistical handling.

Direct handling refers to the direct removal of identified anomalous values. This approach is suitable when the number of anomalies is small and their impact on subsequent analysis is significant.

Statistical handling refers to replacing anomalous values with statistical estimates of the dataset, such as the mean or median. This method is more appropriate when anomalies occur frequently and deleting them may lead to substantial information loss.

The mathematical expression of statistical anomaly processing is shown in Formula (2).。

$$x_i^{'} = \begin{cases} X_\mu \\ median(X) \end{cases} \tag{2}$$

where $X^{'}$ denotes the data value after anomaly replacement (imputation).

2.1.5. Missing Value Localization

Missing value localization refers to identifying the exact positions of missing entries within a dataset (e.g., tables or matrices) and calculating the proportion of missing data.

In this study, an index-based localization method is adopted. For structured data, missing cells can be directly identified using row indices and column names [8–9].

For a dataset *D*, the mathematical expression for missing value localization is given in Formula (3).

$$D_{i,j} = NaN / Null \tag{3}$$

where (i, j) denotes the marked position of a missing value in the dataset.

2.1.6. Missing Value Processing

Missing values can be regarded as a special type of anomaly. Therefore, their treatment follows the same strategy as anomaly processing. The specific handling methods are described in Section 2.1.4.

2.2. Data Integration Module

The data integration module is designed to process multi-source heterogeneous data. Through unified rules and technical approaches, it performs data aggregation, cleaning, transformation,

ISSN: 1813-4890

association, and fusion, ultimately generating high-quality datasets with consistent structure and semantics.

2.2.1. Data Storage Module

Given that the automotive industry involves multiple heterogeneous data sources with different dimensions and formats, the data storage module is responsible for persistent storage of data. A layered storage architecture is adopted.

(1) **Hot data storage:** Redis is employed to cache high-frequency access data, supporting millisecond-level queries;

(2) **Structured data storage:** MySQL and Hive are used to store structured data, supporting complex queries and transactional operations;

(3) **Unstructured data storage:** MongoDB and HDFS are utilized to store images, audio, text, and other unstructured data;

(4) **Data lake storage:** Delta Lake is adopted to construct a data lake for storing both raw and integrated data, supporting data version management and traceability.

2.2.2. Data Fusion Module

Data fusion modules in practical applications can generally be categorized into three levels: data-level fusion, feature-level fusion, and decision-level fusion.

Data-level fusion directly integrates raw data (or cleaned data) from multiple sources while preserving the finest-grained information.

Feature-level fusion first extracts representative features from each data source and then combines multi-dimensional features into a unified feature set.

Decision-level fusion integrates the independent analytical results generated from different data sources to produce a final decision outcome.

In this study, data-level fusion is adopted. This approach performs fine-grained merging of underlying data to retain the maximum amount of original information.

The mathematical formulation of data-level fusion is shown in Formula (4).

$$D_{con} = C(D_1, D_2, \cdots, D_k) = \{(x_{11}, \cdots, x_{1m}, x_{21}, \cdots, x_{2m}, x_{k1}, \cdots, x_{km}, K)\} \quad (4)$$

where Y denotes the fused dataset, $f(\cdot)$ represents the fusion function, and X_k denotes the k-th underlying data source.

2.3. Data Mining

2.3.1. Feature Selection

Feature selection aims to eliminate irrelevant, duplicated, or redundant features, thereby reducing computational cost and mitigating the risk of model overfitting. At the same time, it retains features that are highly correlated with the target variable to improve prediction accuracy and classification performance.

Feature selection can be regarded as a subproblem of model training, where the optimal feature subset is determined by evaluating model performance over different candidate subsets, depending on the specific learning algorithm.

A commonly adopted strategy is Recursive Feature Elimination (RFE), which starts with the full set of features and iteratively removes those that contribute the least to model performance until a predefined number of features is reached.

In this study, a greedy search strategy is employed. The algorithm begins with an empty feature set and progressively adds the feature that yields the largest performance improvement at each step, until no further improvement can be achieved.

ISSN: 1813-4890

2.3.2. Function Selection

Function selection in data mining refers to choosing appropriate analytical techniques according to specific business objectives in order to satisfy targeted analysis or prediction requirements.

Classification:

Data are assigned to different categories based on their features. Core algorithms such as decision trees and random forests are employed to determine whether a sample belongs to a particular class. For example, automotive customers can be segmented into high-value, medium-value, and low-value groups.

(2) Clustering:

Similar data samples are automatically grouped using unsupervised algorithms such as K-Means, hierarchical clustering, and DBSCAN, without predefined labels. For instance, charging users can be divided into high-frequency and low-frequency groups according to their charging behaviors.

Regression:

Regression methods predict continuous numerical values using algorithms such as gradient boosting trees and neural networks.

2.4. Model Validation

The core objective of machine learning model validation is to systematically evaluate the generalization ability of a model and determine whether it can stably and accurately solve practical business problems, while avoiding overfitting or underfitting.

Different validation strategies are selected according to specific task types [11].

(1) Classification/Clustering Model Evaluation

For classification and clustering tasks, Accuracy is adopted as the primary evaluation metric. The calculation of Accuracy is shown in Formula (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Where *TP* denotes True Positives, *TN* denotes True Negatives, *FP* denotes False Positives, and *FN* denotes False Negatives.

(2) Regression Model Evaluation

For regression tasks, Mean Squared Error (MSE) is employed as the evaluation metric. The calculation is given in Formula (6).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

Where y_i represents the ground-truth value and $\hat{y}_i$ denotes the predicted value.

3. Analysis and Prospects

The future development of the machine learning-based automotive data integration system will focus on achieving more comprehensive data coverage, more real-time processing capabilities, enhanced security and reliability, more intelligent decision-making, and broader application scenarios.

To this end, further breakthroughs are required in key technologies such as multimodal data fusion, edge intelligence, and privacy-preserving computing, while simultaneously addressing industry compliance and standardization requirements.

Ultimately, the system is expected to evolve from a conventional data integration tool into a core platform that supports the intelligent transformation of the automotive industry. By

ISSN: 1813-4890

providing robust data foundations and technical guarantees, it will facilitate the advancement of autonomous driving, vehicle–road collaboration, and smart transportation systems.

References

[1] J. Guo, Z. Wang, Y. Guo, *et al.*, "Anomaly behavior detection method for charging pile services based on multi-source data fusion," *Electronic Design Engineering*, vol. 34, no. 3, pp. 94–98, 2026. DOI: 10.14022/j.issn1674-6236.2026.03.020.

[2] F. Zhang, "Data fusion and decision support technologies in intelligent transportation systems," *Scientific* and *Technological Innovation*, no. 3, pp. 93–96, 2026.

[3] C. Bi, Z. Zhang, L. Cao, *et al.*, "Construction of a multi-source data fusion-driven intelligence perception system for future industries," *Journal of Modern Information*, pp. 1–14, 2026. [Online]. Available: https://link.cnki.net/urlid/22.1182.g3.20260202.0954.002.

[4] Z. Gao, "Research on time-series data classification and cleaning based on an improved clustering algorithm," *Automation & Instrumentation*, no. 1, pp. 35–39, 2026. DOI: 10.14016/j.cnki.1001-9227.2026.01.035.

[5] M. Shan, S. Yu, Z. Yan, *et al.*, "Massive heterogeneous data cleaning method based on an improved LSTM neural network," *Technology and Market*, vol. 33, no. 1, pp. 50–53, 58, 2026.

[6] T. Hou, "Identification and processing methods of abnormal values in hydraulic engineering monitoring data," *Innovation in Science and Technology*, vol. 15, no. 35, pp. 148–151, 2025. DOI: 10.19981/j.CN23-1581/G3.2025.35.034.

[7] S. Zheng, D. Hai, Y. Wu, *et al.*, "Outlier detection in sensor network linear data based on the K-Medoids algorithm," *Journal of Dongguan University of Technology*, vol. 32, no. 5, pp. 33–38, 132, 2025. DOI: 10.16002/j.cnki.10090312.2025.05.009.

[8] Y. Li, T. Wang, B. Pang, "Comparative study of missing value imputation methods from an interpretability perspective," *Computer Science*, vol. 52, suppl. 2, pp. 614–621, 2025.

[9] J. Geng, G. Tong, C. Wang, *et al.*, "Missing value imputation for dam deformation monitoring data based on random forest and CNN-GRU models," *Water Power*, pp. 1–10, 2026. [Online]. Available: https://link.cnki.net/urlid/11.1845.tv.20251113.1515.007.

[10] D. Hu and J. Yang, "IP positioning and mapping method based on multi-source data fusion and dynamic clustering," *Journal of Information Security Research*, vol. 12, no. 2, pp. 164–173, 2026.

[11] M. Cao, S. Yin, S. Li, *et al.*, "Risk assessment of floor water inrush under pressure mining based on confusion matrix analysis," *Coal Science and Technology*, vol. 53, no. 9, pp. 407–417, 2025.