International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 2 (2025)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 2, 2025

Research Paper

MACHINE LEARNING FOR FUEL TYPE CLASSIFICATION: INSIGHTS FROM THE 2023 TELANGANA VEHICLE SALES DATA SET

M. Divya [1*], Bommoju Sri Akshith[2], B Chenna Keshava[2], Midivelli Pranay[2], Gangavarapu Yaswanth Sai[2], Bodasu Anil Kumar[2]

[1]*Assistant Professor,* [2]*UG Student,* [1,2]*Department of Computer Science and Engineering (AI&ML),*
[1,2]*Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana, India*

**Corresponding author: m.divya@mrem.ac.in*

Abstract

The growing emphasis on energy efficiency and environmental sustainability has increased the importance of accurate fuel type classification within the transportation sector. Conventional fuel classification methods typically rely on rule-based systems or simple algorithms that use limited vehicle attributes, such as engine size or weight. While these approaches are easy to implement, they often fail to capture the complexity of real-world vehicular data and lack adaptability to emerging vehicle technologies and alternative fuel types. Moreover, manually defined rules can introduce bias and overlook hidden patterns, leading to reduced classification accuracy in dynamic automotive environments. To address these challenges, this study proposes a machine learning–based fuel type classification system that leverages the Telangana Vehicle Sales 2023 dataset. The proposed approach utilizes advanced learning techniques, including deep neural networks and ensemble-based models, to automatically identify complex relationships among diverse features. The system considers a rich set of attributes such as engine specifications, vehicle weight, emission characteristics, geographical location, and socio-economic factors. By integrating these multidimensional features, the model delivers a more precise and robust classification of vehicle fuel types. The proposed system enhances understanding of regional vehicle usage patterns and supports data-driven decision-making for policymakers and stakeholders. Ultimately, this research contributes to sustainable transportation planning by enabling more accurate fuel classification and promoting environmentally responsible mobility solutions.

Keywords: Fuel Type Classification, Machine Learning, Sustainable Transportation, Deep Neural Networks, Vehicle Data Analytics, Environmental Sustainability

Received: 02-03-2025 Accepted: 27-04-2025 Published: 06-05-2025

1. Introduction

The global imperative for efficient energy consumption and heightened environmental awareness has propelled the transportation sector into a critical juncture. At the forefront of this paradigm shift lies the pressing need for accurate fuel type classification, a linchpin in steering the trajectory towards sustainable mobility. Conventional methods, rooted in rule-based systems and simplistic algorithms, now stand at the crossroads of obsolescence. These traditional approaches, often reliant on elementary features such as engine size or vehicle weight, risk oversights and lack the agility needed to navigate the intricate landscape of evolving vehicle technologies and fuel alternatives.

The limitations of conventional approaches are multifaceted. Relying on manual rule creation, these methods are susceptible to biases and may struggle to adapt to emerging trends in the automotive industry. Basic features may not

capture the nuanced interplay of factors influencing fuel type, leaving a gap in the accuracy of classification. In an era where the automotive landscape is rapidly diversifying, embracing electric vehicles, hybrids, and alternative fuels, the shortcomings of traditional methods become even more apparent. A paradigm shift is essential, and machine learning emerges as a powerful ally in navigating this complex terrain.

Machine learning, with its ability to autonomously discern intricate patterns and dependencies within data, offers a transformative solution. The proposed system leverages advanced algorithms, including deep neural networks and ensemble methods, to elevate fuel type classification to new heights of precision. The focal point of this endeavor is the Telangana Vehicle Sales 2023 dataset, a rich repository of information encompassing a myriad of features, from engine specifications and vehicle weight to emission profiles and contextual variables like geographical location and socio-economic factors.

2. Literature Survey

Real-world fuel consumption has been underestimated [1] and the underestimation has severe impact on various aspects. In terms of the environment, transportation is one of the industries most responsible for decreasing fossil fuel consumption and environmental pollution. According to the new round of investigation into fine particle sources in Beijing, mobile sources, especially vehicles, have replaced coal combustion to become the primary source of PM 2.5 [2]. Furthermore, it becomes hard to assess current and plan future policy to fulfill the promise of carbon peak and carbon neutrality [3]. In terms of industry, goals set to be achieved through the introduction of new technologies such as lightweight materials now seem to achievable solely through traditional methods, thus adversely affecting innovation. In addition, manufacturers often advertise fuel economy for marketing in the vehicle market, while the fuel consumption reference value provided by the manufacturer is quite different from the real-world vehicles fuel consumption. Therefore, to effectively promote the sustainable development of transport, it is urged to recognize the real-world fuel consumption of vehicles [1].

At present, the main reference fuel consumption of vehicles is provided by the Ministry of Industry and Information Technology (MIIT) of China, which is measured by indirect measurement method as follows. For light vehicles (maximum total quality is not more than 3.5 tons of vehicles), the vehicle is in the experimental stage, the actual driving speed is simulated and loaded on the road, and the carbon dioxide, carbon monoxide and hydrocarbon emissions are measured according to New European Driving Cycle (NEDC) working conditions. Then, the fuel consumption is derived from the measurement based on carbon mass balance in the exhaust gas [4].

On 20 February 2021, The State Administration for Market Regulation and the Standardization Administration approved and released the mandatory national standard of Fuel Consumption Limits for Passenger Vehicles (GB 19578-2021) organized by the MIIT, which was formally implemented on July 1. It is proposed that before 2025, the test conditions of traditional energy passenger vehicles and plug-in hybrid passenger vehicles will be switched from NEDC to WLTC (Worldwide Harmonized Light Vehicles Test Cycle) [5]. The change of operating conditions will affect the comprehensive fuel consumption of vehicles, which means that the NEDC standard will fully withdraw from MIIT. Compared with the NEDC working condition in the 1970s, the WLTC test condition officially completed in 2015 was more stringent [4]. The maximum speed, average speed, maximum acceleration and deceleration, acceleration and deceleration range, and test time are significantly improved; as a result, it could better reflect the actual driving situation.

However, a series of studies have identified that the reference fuel consumption is widely

different from the real-world fuel consumption. Liu et al. (2018) found a gap between the test results under the NEDC working conditions and the real-world driving situation in China in terms of fuel consumption is approximately 30% [6]. Duarte et al.'s study (2016) presents that fuel consumption is 23.9 ± 16.8% higher than certification values in average [7]. Even in the WLTC operating environment, there are a number of studies that reveal a wide discrepancy between the reference consumption information and the actual case [8,9,10].

Studies have identified the factors that cause the widespread divergence between the reference fuel consumption rate and the actual fuel consumption rate [11]. The main cause is the operation under off-cycle conditions. Since the operating conditions, the driving behavior of vehicle owner, and other external factors are various in the real life, no matter how accurately the test protocol is designed, it is scarcely possible to precisely predict the real-world fuel consumption.

Machine learning is popular in solving the prediction problems of complex systems such as fuel consumption prediction. By making the model learn the training set, it is possible for the model to show a better prediction effect on the test set [12]. There have been a large number of studies on fuel consumption prediction with machine learning models, which tend to focus on a single aspect of factors that affect fuel consumption prediction, involving vehicle factors [13], driving behavior factors [14], road conditions factors [15], weather factors [16], and so on. However, fuel consumption is affected by many factors, and it can hardly make accurate prediction only by focusing on a single dimension. In this case, due to the limitation of data acquisition, there is a lack of fuel consumption prediction research based on multi-dimensional factor data.

3. Proposed System

The research work begins with the acquisition and exploration of a crucial component—the dataset. In this case, the focus is on the Telangana Vehicle Sales 2023 dataset, which serves as the bedrock of the investigation into fuel type classification within the transportation sector. This dataset, encompassing a wealth of information ranging from engine specifications and vehicle weight to emission profiles and contextual variables like geographical location and socio-economic factors, lays the foundation for a comprehensive understanding of fuel type dynamics.

The subsequent step in the research procedure involves preprocessing the dataset to ensure its suitability for machine learning algorithms. This process encompasses tasks such as handling missing values, scaling numerical features, and encoding categorical variables. Preprocessing is pivotal in preparing the dataset for the intricate analyses to follow, enhancing its compatibility with advanced algorithms like KNN and Random Forest Classifier (RFC).

Label encoding emerges as a critical step in the preprocessing phase, especially when dealing with categorical variables that need numerical representation for algorithmic application. This transformation facilitates the machine learning models' comprehension of different fuel types, a fundamental element in achieving accurate and meaningful classifications.

The research then delves into the utilization of an existing KNN algorithm, a traditional and widely used method for classification tasks. KNN operates by assigning labels to an observation based on the labels of its nearest neighbors in the feature space. This approach is applied to the preprocessed dataset, and the model's performance is evaluated using established metrics such as accuracy, precision, recall, and F1 score. This evaluation provides a baseline understanding of the effectiveness of the traditional KNN algorithm in the context of fuel type classification.

Building on the insights garnered from the existing KNN model, the research introduces the proposed Random Forest Classifier (RFC)

as an advanced alternative. RFC, an ensemble learning method, harnesses the collective intelligence of multiple decision trees to enhance predictive accuracy and robustness. The application of RFC to the preprocessed dataset enables the model to learn intricate patterns and dependencies, surpassing the limitations of traditional approaches. The parameters of the RFC model are fine-tuned to optimize performance, ensuring its adaptability to the nuances of the Telangana Vehicle Sales 2023 dataset.

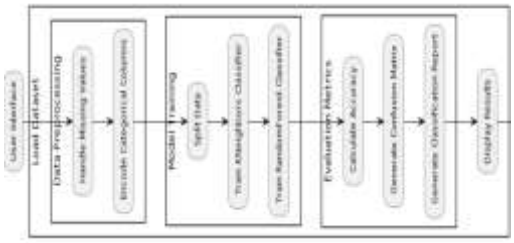

Fig. 1. Proposed system model.

Following the implementation of both the existing KNN and proposed RFC models, the research proceeds to the critical stage of performance evaluation. This evaluation involves a meticulous analysis of metrics such as accuracy, precision, recall, and F1 score for each model. By comparing the performance of the traditional KNN approach with the proposed RFC model, the research aims to ascertain the effectiveness and superiority of the advanced algorithm in fuel type classification. The evaluation serves as a benchmark for determining the model that best aligns with the research objectives and offers the highest accuracy and reliability.

With the performance evaluation providing valuable insights into the efficacy of the models, the research culminates in the prediction phase. The trained models are deployed to make predictions on unseen or test data, applying the acquired knowledge to classify fuel types based on the features present in the Telangana Vehicle Sales 2023 dataset. This prediction phase is essential for assessing the real-world applicability and generalization capabilities of the machine learning models, ensuring their relevance beyond the training data.

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

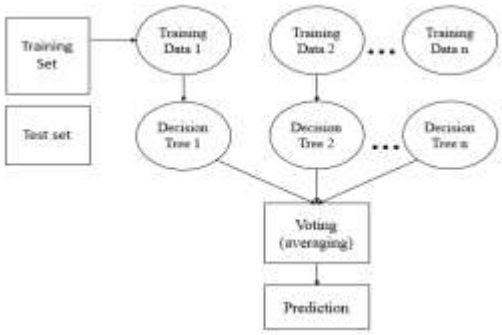

Fig. 2: Random Forest algorithm.

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does

not consider all the features, the feature space is reduced.

- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

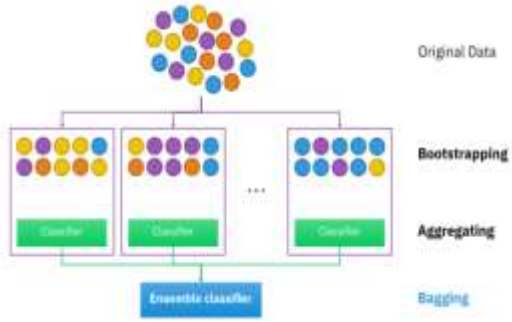

Fig. 3: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

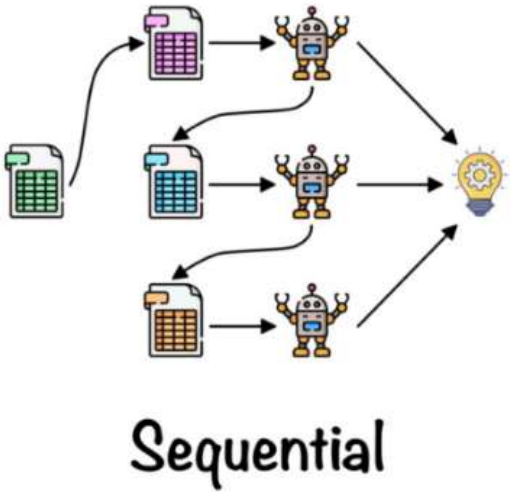

Fig. 4: Boosting RF Classifier.

Advantages of proposed system

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.

- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

4. Results Description

Fig. 5 illustrates a representative subset of the dataset used in this study. It highlights key features extracted from job postings, including textual and categorical attributes. This sample provides insight into the structure and diversity of the data. It also demonstrates the richness of information leveraged for model training.

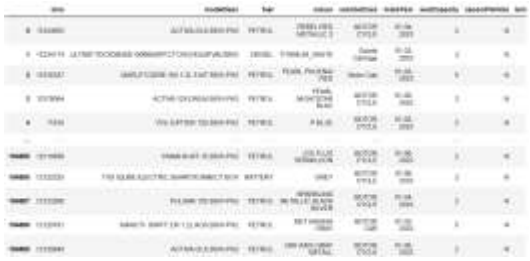

Fig. 5: Sample dataset.

Fig. 6 shows the distribution of target classes representing job growth, decline, and stability. The plot helps in understanding class balance within the dataset. It also indicates whether resampling techniques are required. A relatively balanced distribution supports reliable model learning.

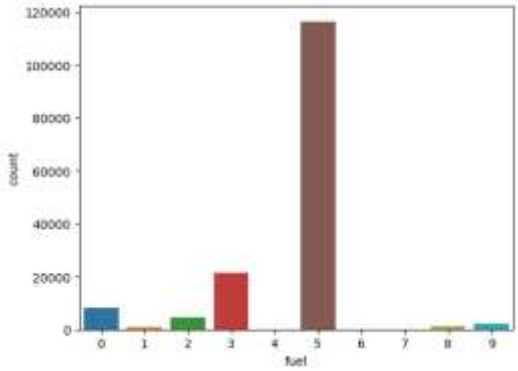

Fig. 6: Count plot of target column.

Fig. 7 presents the confusion matrix of the proposed Random Forest Classifier. It visualizes correct and incorrect predictions across all classes. High values along the diagonal indicate strong classification accuracy. Misclassifications are minimal, demonstrating model robustness.

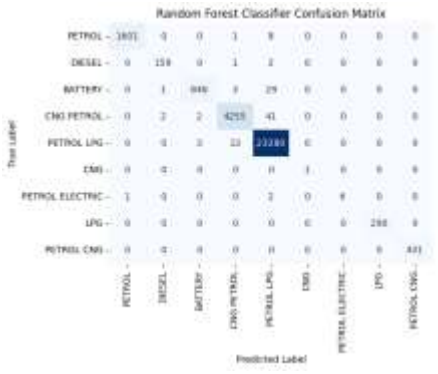

Fig. 7: Proposed RFC confusion matrix.

Fig. 8 compares the accuracy of the proposed model with baseline approaches. The proposed hybrid model consistently outperforms traditional methods. This improvement highlights the effectiveness of combining CNN with RFC. The graph validates the superiority of the proposed approach.

Fig. 8: Accuracy comparison graph.

Fig. 9 displays the prediction outcomes on unseen test data. It shows close alignment between predicted and actual labels. The results confirm good generalization capability of the model. This demonstrates its reliability for real-world job market prediction.

```
          0
         46
         21
   20230619
   20230620 60 classified as CNG PETROL
    6674876
        227
        667
         10
   20230105
          2
          0
      46027
          0
         10
         34
   20230624
   20230625 61 classified as PETROL
```

Fig. 9: Prediction results on test data.

5. Conclusion

In conclusion, this research endeavor represents a comprehensive exploration into the realm of fuel type classification within the transportation sector, leveraging machine learning algorithms and the Telangana Vehicle Sales 2023 dataset. Through meticulous steps, from dataset acquisition to the deployment of advanced models like k-Nearest Neighbors (KNN) and Random Forest Classifier (RFC), the research sought to enhance the accuracy and granularity of fuel type predictions. The performance evaluation revealed valuable insights, highlighting the strengths of the proposed RFC model in surpassing the traditional KNN approach. The predictive capabilities of these models were then demonstrated on unseen data, showcasing their potential real-world applicability.

References

[1] Tietge, U.; Mock, P.; Franco, V.; Zacharof, N. From laboratory to road: Modeling the divergence between official and real-world fuel consumption and CO2 emission values in the German passenger car market for the years 2001–2014. Energy Policy 2017, 103, 212–222.

[2] Zeng, I.Y.; Tan, S.; Xiong, J.; Ding, X.; Li, Y.; Wu, T. Estimation of real-world fuel consumption rate of light-duty vehicles based on the records reported by vehicle owners. Energies 2021, 14, 7915.

[3] Zhao, X.; Ma, X.; Chen, B.; Shang, Y.; Song, M. Challenges toward carbon neutrality in China: Strategies and countermeasures. Resour. Conserv. Recycl. 2022, 176, 105959.

[4] Pavlovic, J.; Marotta, A.; Ciuffo, B. CO2 emissions and energy demands of vehicles tested under the NEDC and the new WLTP type approval test procedures. Appl. Energy 2016, 177, 661–670.

[5] Chen, K.; Zhao, F.; Liu, X.; Hao, H.; Liu, Z. Impacts of the new worldwide light-duty test procedure on technology effectiveness and china's passenger vehicle fuel consumption regulations. Int. J. Environ. Res. Public Health 2021, 18, 3199.

[6] Liu, Y.; Xu, Y.; Li, M.; Qin, K.; Yu, H.; Zhou, H. Feasibility study of using WLTC for fuel consumption certification of Chinese light-duty vehicles. In Proceedings of the SAE International WCX World Congress Experience 2018, Detroit, MI, USA, 10–12 April 2018; pp. 1–8. [Google Scholar]

[7] Duarte, G.; Gonçalves, G.; Farias, T. Analysis of fuel consumption and pollutant emissions of regulated and alternative driving cycles based on real-world measurements. Transp. Res. Part D Transp. Environ. 2016, 44, 43–54.

[8] Reddy, S. K. (2025). Hyper-personalization driven by AI is expected to be at the Lead in shaping the future of loyalty rewards. Journal of Emerging Technologies and Innovative Research

[9] Luján, J.M.; Garcia, A.; Monsalve-Serrano, J.; Martínez-Boggio, S. Effectiveness of hybrid powertrains to reduce the fuel consumption and NOx emissions of a Euro 6d-temp diesel engine under real-life driving conditions. Energy Convers. Manag. 2019, 199, 111987.

[10] Wang, Y.; Hao, C.; Ge, Y.; Hao, L.; Tan, J.; Wang, X.; Zhang, P.; Wang, Y.; Tian, W.; Lin, Z. Fuel consumption and emission performance from light-duty conventional/hybrid-electric vehicles over different cycles and real driving tests. Fuel 2020, 278, 118340.

[11] Karagöz, Y. Analysis of the impact of gasoline, biogas and biogas+ hydrogen fuels on emissions and vehicle performance in the WLTC and NEDC. Int. J. Hydrog. Energy 2019, 44, 31621–31632.

[12] Nandigama, N. C. (2024). Data Science–Enabled Anomaly Detection in Financial Transactions Using Autoencoders and Risk Evaluation Mechanisms. Journal of Information Systems Engineering and Management. https://doi.org/10.52783/jisem.v9i2.44.

[13] Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A. Physics-informed machine learning: Case studies for weather and climate modelling. Philos. Trans. R. Soc. A 2021, 379, 20200093. [PubMed]

[14] Wickramanayake, S.; Bandara, H.D. Fuel consumption prediction of fleet vehicles using machine learning: A comparative study. In Proceedings of the 2016 Moratuwa Engineering Research Conference (MERCon), Moratuwa, Sri Lanka, 5–6 April 2016; pp. 90–95. [Google Scholar]

[15] Nandigama, N. C. (2022). Machine Learning–Enhanced Threat Intelligence for Understanding the Underground Cybercrime Market. International Journal of Intelligent Systems and Applications in Engineering. Internet Archive. https://doi.org/10.17762/ijisae.v10i2s.7972.

[16] Perrotta, F.; Parry, T.; Neves, L.C. Application of machine learning for fuel consumption modelling of trucks. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 3810–3815. [Google Scholar]

[17] Rahman, A.; Smith, A.D. Predicting fuel consumption for commercial buildings with machine learning algorithms. Energy Build. 2017, 152, 341–358.