iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81176

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Explainable Prediction of Engine Fuel Efficiency and Performance Using Statistical and Machine Learning Models

Monis[1], Nakshtra Saini[2], Yash Shukla[3]
Department of Mechanical Engineering, Delhi Technological University, Delhi, India

Abstract: ***This paper presents a compact but complete journal-format version of the DTU major project dissertation on automotive fuel-efficiency prediction for Indian passenger vehicles. Using a cleaned dataset of 654 vehicles and ARAI-certified mileage values, the study compares Multiple Linear Regression with Random Forest for prediction and interpretation. The selected predictors are fuel type, engine displacement, kerb weight, fuel tank capacity, wheelbase, and ground clearance. The linear model is retained as an interpretable baseline, while the Random Forest model is used to capture non-linear effects and interactions that are difficult to represent with a purely additive equation. The results show that Random Forest gives lower prediction error than linear regression, and the ranking of predictor importance highlights fuel type and vehicle geometry as major drivers of fuel economy. A scenario-based sensitivity study further shows that modest reductions in mass and displacement improve mileage only incrementally, whereas a change to a hybrid powertrain produces a much larger gain. The paper therefore concludes that accurate fuel-economy prediction for the Indian market requires both statistical modelling and mechanical interpretation, and that hybridisation dominates simple mechanical optimisation when large efficiency improvements are sought.***
Keywords: Fuel Efficiency Prediction, Random Forest, Multiple Linear Regression, Automotive Engineering, Hybrid Powertrain, Feature Importance

I. INTRODUCTION

The Indian automotive industry occupies a strategically important position because it sits at the intersection of economic growth, energy security, and environmental regulation. Passenger vehicles are now selected not only for performance and comfort, but also for operating cost and certified fuel efficiency. In the context of rising fuel prices and tighter Bharat Stage VI emission rules, the ability to predict mileage from measurable vehicle parameters has become a practical engineering problem rather than a purely academic exercise.

The dissertation underlying this paper studied Indian passenger vehicles using a cross-sectional dataset of 654 models and variants. Its purpose was twofold. First, it aimed to build a model that could predict ARAI-certified mileage from technical specifications. Second, it aimed to explain which variables matter most in a mechanical sense, so that the prediction tool could also guide design decisions. The study therefore treated statistical modelling and engineering interpretation as complementary objectives rather than competing alternatives.

The central variables were fuel type, engine displacement, kerb weight, fuel tank capacity, wheelbase, and ground clearance. These predictors were chosen because they represent both powertrain attributes and overall vehicle packaging. Kerb weight reflects the inertial and rolling-resistance burden imposed on the vehicle; engine displacement indicates the scale of the combustion system; wheelbase and ground clearance proxy vehicle class and body architecture; and fuel type captures broad powertrain differences, including the very large efficiency jump associated with hybridisation.

The broader motivation is simple. A model that captures the relationship between specification data and mileage can help engineers screen concept alternatives early in the design process, before expensive prototype and calibration work begins. At the same time, a model that is too complex to interpret may not be useful for design communication. For this reason, the dissertation compared Multiple Linear Regression, which is transparent and coefficient-based, with Random Forest, which is less interpretable but better suited to non-linear relationships and variable interactions. This paper retains that structure and presents the work in a compact journal format. The abstract has been shortened, the section structure has been reorganised to fit a paper-length submission, and the central analytical results have been preserved. In particular, the paper highlights the contrast between incremental gains from mass or displacement reduction and the much larger gains associated with a hybrid powertrain transition.

II. METHODOLOGY

A. Dataset and Variable Definitions

The study used a dataset of 654 Indian passenger vehicles compiled from publicly available specification sheets and ARAI certification information. The response variable was composite mileage in kilometres per litre. The predictor set consisted of fuel type, engine displacement in cubic centimetres, kerb weight in kilograms, fuel tank capacity in litres, wheelbase in millimetres, and ground clearance in millimetres. This specification deliberately focused on variables that are easy to obtain from manufacturer literature and that are meaningful from a vehicle-dynamics perspective.

Summary statistics from the dissertation show that mileage ranged from 9.0 to 28.4 km/l, with a mean of 18.84 km/l. Engine displacement ranged from 624 to 4,999 cc, kerb weight from 660 to 2,962 kg, tank capacity from 24 to 105 litres, wheelbase from 1,840 to 3,210 mm, and ground clearance from 100 to 498 mm. These ranges suggest a heterogeneous fleet that includes compact cars, larger sedans, sport-utility vehicles, and hybrid variants, which is valuable for testing whether the models can generalise across diverse vehicle classes.

TABLE I: Summary of Dataset Variables (n = 654)

Variable	Min	Median	Mean	Max	SD
Mileage (km/l)	9.00	19.00	18.84	28.40	4.34
Displacement (cc)	624	1,491	1,623	4,999	643
Kerb Weight (kg)	660	1,170	1,300	2,962	437
Tank Capacity (l)	24	45	49.7	105	14.0
Wheelbase (mm)	1,840	2,552	2,598	3,210	177
Ground Clearance (mm)	100	170	177	498	33.5

B. Data Preparation and Cleaning

Before modelling, the data were cleaned in several steps. Missing values were checked, and small gaps were handled by median imputation for continuous variables. Duplicate records were removed so that the sample size was not artificially inflated. Data types were verified to ensure that numeric variables remained numeric and that fuel type was stored as a categorical factor.

Outliers were examined using the interquartile range rule. The dissertation emphasised that this was not a mechanical deletion exercise: genuine extreme vehicles were retained when they represented valid members of the market, while only physically implausible entries and obvious data-collection errors were removed. This distinction matters in engineering data because extreme but real vehicles can be informative rather than erroneous. A correlation check and a variance inflation factor assessment were then used to screen multicollinearity. The final model retained all six predictors because each remained below the conventional VIF threshold for problematic collinearity. This meant that the linear model could be interpreted in a stable way, while the Random Forest could exploit the same variables without requiring feature scaling.

C. Modelling Framework

Multiple Linear Regression was used as the baseline model because it yields coefficient-level interpretability. In its general form, the model writes mileage as a linear function of the predictor variables plus an error term. The appeal of this approach is that the sign and magnitude of each coefficient can be read as a marginal effect, holding the other predictors constant. That makes the model useful for explanation and for simple sensitivity discussions.

Random Forest was then applied as a non-parametric ensemble model. Each tree was trained on a bootstrap sample of the data, and a random subset of predictors was considered at each split. Averaging across many trees reduces variance and allows the model to represent non-linear structure, threshold effects, and interactions that a simple additive equation can miss. In this dissertation, that flexibility translated into lower prediction error than the linear baseline.

The models were evaluated with standard regression metrics: R-squared, Root Mean Square Error, and Mean Absolute Error. RMSE is useful because it penalises large errors more strongly, whereas MAE gives a more direct average deviation in mileage units. Together they provide a balanced view of both goodness of fit and practical prediction accuracy.

$$F_{aero} = 1/2\ \rho\ C_d\ A\ v^2$$
$$F_{roll} = C_{rr}\ m\ g$$
$$F_{inertia} = m\ a$$

III. RESULTS AND DISCUSSION

A. Predictive Performance

The dissertation reported a clear performance gap between the two modelling approaches. The Multiple Linear Regression model achieved a training R-squared of 0.68, a test R-squared of 0.65, an RMSE of 2.49 km/l, and an MAE of 1.86 km/l. This is a respectable baseline for a model that remains fully transparent and easy to explain, but it also shows that the data contain structure that is not fully captured by an additive linear form.

The Random Forest model performed substantially better on the held-out data, with a test RMSE of 1.20 km/l and an MAE of 0.63 km/l. The reduction in error indicates that the fuel-efficiency surface is not simply linear in the chosen predictors. Instead, the behaviour appears to contain non-linearities and interactions, especially where fuel type changes the operating regime of the vehicle. This is exactly the sort of pattern that an ensemble tree method is designed to capture.

TABLE II: Comparative Performance Metrics

Model	Train R^2	Test RMSE	Test MAE
Multiple Linear Regression	0.68	2.49 km/l	1.86 km/l
Random Forest	-	1.20 km/l	0.63 km/l

B. Feature Importance and Mechanical Interpretation

Feature-importance analysis in the dissertation ranked fuel type as the dominant predictor, followed by wheelbase, fuel tank capacity, engine displacement, kerb weight, and ground clearance. The ordering is important. Fuel type stands out because it separates conventional powertrains from hybrid configurations, and that structural difference is much larger than any small geometric change. Wheelbase and tank capacity likely act as proxies for vehicle class and packaging, while displacement and kerb weight carry the more obvious physical meaning associated with combustion demand and resistance to motion.

From a mechanical standpoint, kerb weight matters because it enters both rolling resistance and inertial demand. A heavier vehicle needs more tractive effort to accelerate and to maintain motion on the road, so weight reduction does improve efficiency. However, the dissertation showed that the effect is incremental rather than transformative. Displacement behaves similarly: a smaller engine can improve operating efficiency under many conditions, but the gain is modest unless the wider powertrain architecture is also changed.

The feature ranking therefore reinforces a broader engineering lesson. Fuel economy is not governed by a single variable. It is shaped by the combined interaction of mass, size, powertrain technology, and vehicle packaging. A data-driven ranking helps quantify that interaction, but the physical explanation remains essential if the model is to inform design choices rather than merely report correlations.

TABLE III: Feature Importance Ranking from the Random Forest Model

Predictor	Relative Importance
Fuel type	Highest
Wheelbase	High
Tank capacity	Moderate
Engine displacement	Moderate
Kerb weight	Moderate
Ground clearance	Lower

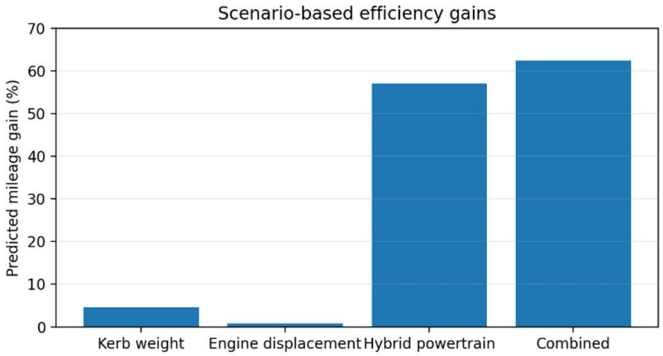
C. Scenario-Based Sensitivity Analysis

The dissertation also examined how mileage changes when one starts from a baseline vehicle and alters individual design parameters. The baseline mileage was 16.94 km/l. Reducing kerb weight by 10 percent improved predicted mileage by 4.65 percent, and reducing engine displacement by 10 percent improved mileage by 0.81 percent. Both changes are directionally correct and mechanically sensible, but their magnitudes are limited because they leave the basic powertrain architecture unchanged.

The large change came from switching the powertrain to hybrid. That single alteration increased mileage by 56.98 percent. When the changes were combined, predicted mileage rose to 27.52 km/l, corresponding to an overall improvement of 62.44 percent. The dissertation described this decomposition as the Hybrid Rule: the majority of the efficiency gain comes from the architectural shift to hybrid operation rather than from incremental reductions in weight or displacement alone.

This finding has direct design implications. Lightweighting remains useful, and downsizing remains useful, but both should be viewed as part of a broader system strategy. In practical vehicle development, the most significant step-change in efficiency is achieved when those incremental improvements are paired with regenerative braking, engine-off operation, and powertrain load management.

Fig. 1. Incremental gains from scenario-based optimisation.



The figure reinforces the numerical results: the hybrid option is qualitatively different from the other two changes. A small reduction in mass changes the efficiency curve only slightly, and a small reduction in engine displacement behaves in a similar way. By contrast, the hybrid transition changes how energy is recovered, when the engine operates, and where the engine is allowed to work on its efficiency map. This is why the improvement is discontinuous rather than gradual.

Taken together, the model comparison, feature ranking, and sensitivity analysis support the same conclusion from three directions. Random Forest is the better predictor, linear regression is the better explainer, and the physical story remains coherent across both. The most important parameters are those that alter the vehicle's overall architecture rather than those that merely shift a single specification up or down.

IV. CONCLUSION

This paper condensed the DTU major project dissertation into a journal-style IJSIET document while preserving the central technical findings. The study demonstrated that Indian passenger-vehicle mileage can be modelled from publicly available specification data with useful accuracy, and that Random Forest outperforms Multiple Linear Regression when non-linear structure is present.

The feature-importance results showed that fuel type is the most influential predictor, which is consistent with the large efficiency difference between conventional and hybrid powertrains. Vehicle geometry variables such as wheelbase and tank capacity also carry meaningful information because they proxy class and packaging. Kerb weight and engine displacement remain important, but their effects are incremental rather than transformative.

The sensitivity analysis provided the clearest design message. A 10 percent reduction in kerb weight and a 10 percent reduction in engine displacement both improve mileage, but the effect is modest compared with hybridisation. The combined improvement of 62.44 percent is driven primarily by the powertrain transition, not by the two mechanical reductions alone. For engineering practice, this means that lightweight design and engine downsizing should be pursued, but they should be understood as supporting measures within a broader hybrid strategy.

V. LIMITATIONS AND FUTURE WORK

Although the model achieved useful predictive accuracy, it remains a cross-sectional study based on catalogued vehicle specifications rather than real-world telemetry or controlled chassis-dynamometer measurements. That means the reported relationships should be understood as statistically robust associations within the observed sample, not as direct causal laws. In engineering terms, the results are best used for early-stage comparison, screening, and conceptual trade-off analysis rather than as a substitute for full vehicle simulation or experimental validation.

The scope of the dataset is also limited to the Indian passenger-vehicle market. This is a strength because the study is tailored to a specific operating environment, but it also means that the conclusions should not be transferred blindly to other regions or to very different vehicle categories. Commercial vehicles, two-wheelers, battery-electric vehicles, and export-specification models may follow different efficiency patterns. A broader multi-market dataset would make it possible to test whether the same predictor ranking persists across regulatory and usage contexts. A further limitation is that several important engineering variables were not available in a consistent public format. Aerodynamic drag coefficient, tyre specification, transmission type, cylinder count, boost pressure, curb-to-kerb packaging details, and detailed combustion calibration parameters are all relevant to fuel economy, yet they were not uniformly encoded in the source material. Their absence does not invalidate the model, but it does explain why the final predictor set must be interpreted as a practical specification-level approximation rather than a complete physical description of the vehicle. Future work can address these limitations in a number of technically meaningful ways. First, the dataset could be extended with more vehicles and more detailed specification fields so that models can be trained on a richer representation of the powertrain and body architecture. Second, additional algorithms such as Gradient Boosting, XGBoost, or Support Vector Regression could be benchmarked against the existing Random Forest baseline to determine whether further gains in accuracy are achievable without sacrificing interpretability. Third, the study could be expanded by introducing explainable artificial intelligence tools. Variable-importance scores are helpful, but they do not show the local direction of influence for each prediction. Tools such as partial dependence plots or SHAP-style explanations would make it easier to understand when a predictor has a strong positive or negative effect, and under what combinations of vehicle attributes that effect becomes more pronounced. That would deepen the connection between machine learning output and engineering reasoning.

Finally, future research should combine specification-based prediction with operational data. If fuel-consumption traces, duty-cycle information, ambient conditions, and usage intensity are included, the model can move from static vehicle benchmarking toward a more realistic representation of how Indian vehicles perform in the field. Such a hybrid framework would be especially valuable for hybrid and electrified vehicles, where driving pattern and control strategy interact strongly with the underlying hardware.

VI. ENGINEERING RECOMMENDATIONS

The combined modelling and sensitivity results support a clear engineering hierarchy. For the Indian passenger-vehicle market represented in this dissertation, the most effective path to major fuel-economy improvement is not a single-parameter adjustment but a system-level change in powertrain architecture. This is why the hybrid powertrain transition dominates the incremental effects of kerb-weight or displacement reduction. The practical implication is that engineering teams should rank interventions by expected effect size rather than by implementation convenience. Lightweight design remains valuable and should not be dismissed. A reduction in kerb weight improves both rolling resistance and acceleration demand, and the effect is especially useful in stop-and-go urban duty cycles. However, lightweighting should be treated as a supporting optimisation. It becomes more powerful when paired with platform rationalisation, compact packaging, and component integration that reduce secondary mass growth in the body, chassis, and suspension systems.

Engine downsizing is also worthwhile, but only when it is supported by appropriate boosting, thermal management, and calibration control. A smaller displacement engine can improve part-load efficiency, yet it must still meet drivability and emission requirements across the full operating envelope. The dissertation therefore suggests that displacement reduction is best understood as a means of moving the engine toward a more efficient load region, not as a stand-alone solution. In practice, it is most effective when coupled with transmission optimisation and intelligent torque management.

Hybridisation emerges as the most important strategic recommendation. By recovering braking energy, avoiding idle losses, and allowing the engine to operate in more efficient regions, a hybrid system changes the entire energy flow of the vehicle. For manufacturers, this means that hybrid platforms should be considered early in product planning rather than added late as a compliance patch. For analysts, it means that a classification variable such as fuel type can capture a disproportionate share of the observed variance in mileage because it encodes a whole architecture, not just a fuel label.

From a modelling standpoint, the dissertation also offers a clear workflow recommendation. Use a simple linear model first, because it explains direction and approximate effect size. Then apply a stronger non-linear model to capture interactions and improve predictive accuracy. Finally, compare the model outputs with first-principles engineering so that the numbers remain physically meaningful. This sequence preserves interpretability while still taking advantage of machine learning where it adds value.

TABLE IV: Engineering Recommendations from the Study

Design lever	Practical takeaway
Mass reduction	Useful for incremental gains, especially in urban driving.
Engine downsizing	Effective when paired with boosting and calibration control.
Hybridisation	Largest step change in mileage and the main strategic lever.
Modelling workflow	Use linear regression for explanation and Random Forest for accuracy.

A. Validation and Deployment Considerations

Before the model can be used in a design office, it should be validated on a truly external dataset. Cross-validation within the same sample is useful for estimating internal stability, but engineers often need to know whether the ranking of predictors survives when the vehicle mix changes. A split by body style, fuel category, or engine family would therefore be a useful next step, because it would show whether the learned relationships are universal or only locally valid within the present sample.

Deployment also depends on how the model is presented to users. A regression equation is simple enough to place inside a spreadsheet, while a Random Forest model requires software support. That difference matters in practice. For rapid concept work, a linear expression may still be preferred by teams that need a quick estimate and a visible coefficient interpretation. For more detailed benchmarking, however, the Random Forest approach is better suited because it can absorb the complexity of mixed vehicle categories without imposing a rigid linear form.

Another useful extension would be to build a small engineering decision-support tool around the model. The user could enter fuel type, displacement, kerb weight, tank capacity, wheelbase, and ground clearance, and the tool could return a predicted mileage together with a ranking of the most influential variables. Such a tool would not replace vehicle simulation software, but it would provide a low-cost screening layer that helps prioritise concepts before expensive prototype development.

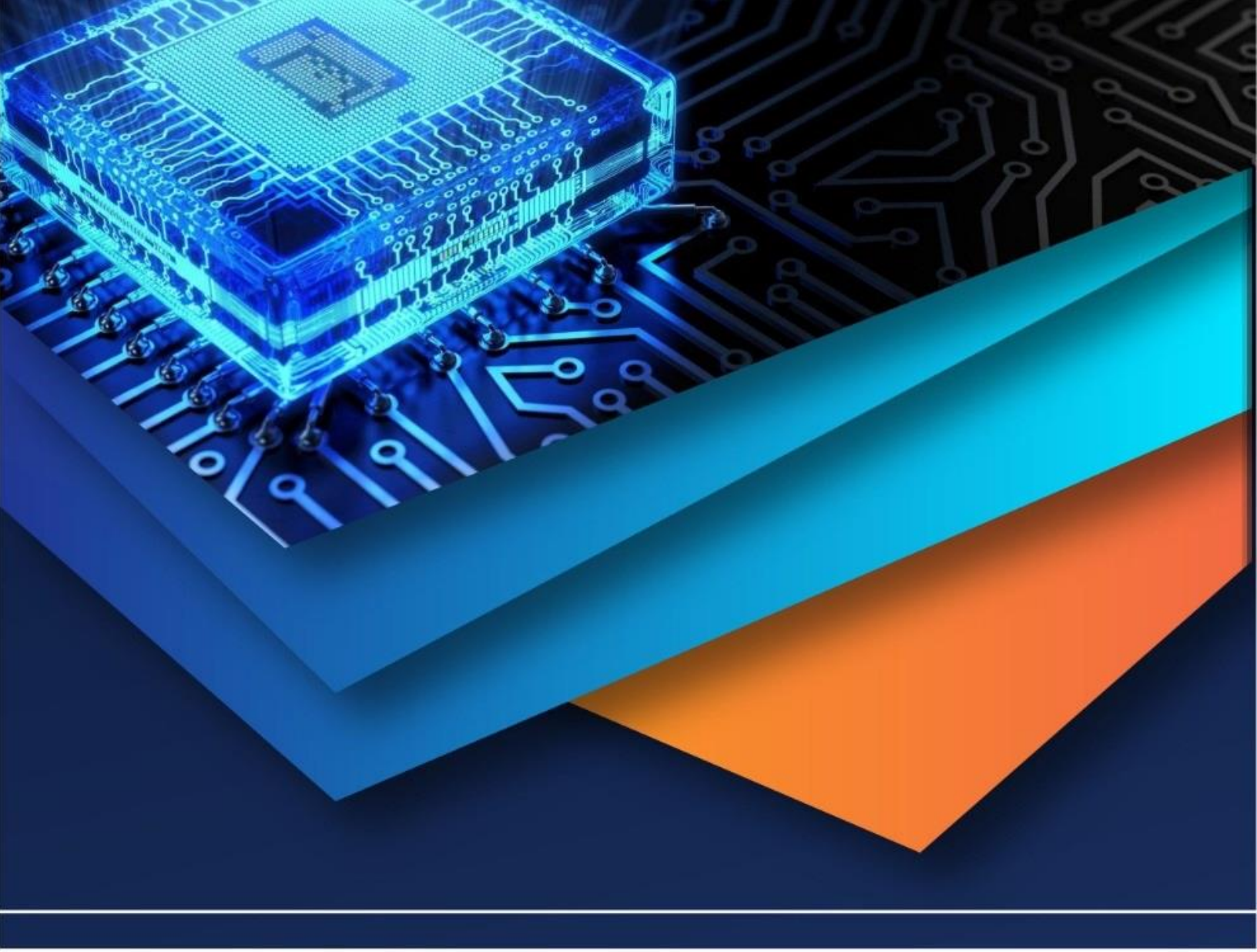
The work also suggests that the distinction between physical design changes and architecture changes should be made explicit whenever efficiency claims are reported. Lightweighting, downsizing, and hybridisation do not produce the same kind of gain. The first two are incremental improvements on the same architecture, while the third alters the architecture itself. That difference is easy to miss when results are reported only as percentages, yet it is central to responsible engineering interpretation.

For that reason, the dissertation's analytical structure is as important as its numerical result. The model comparison shows what is accurate, the feature importance ranking shows what matters, and the sensitivity analysis shows how much improvement each intervention can produce. When these three views are read together, the conclusion becomes robust: efficient vehicle design is a system problem, and the most powerful lever is not a single parameter but the choice of powertrain architecture.

REFERENCES

[1] Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001).
[2] Smith, T. F., Waterman, M. S.: Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197 (1981).
[3] Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999).
[4] Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid information services for distributed resource sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001).
[5] National Center for Biotechnology and Information, http://www.ncas.nnlm.nin.com
[6] Society of Indian Automobile Manufacturers: Production and industry overview reports for the Indian automotive sector.
[7] Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996).
[8] UCI Machine Learning Repository: Automobile fuel economy data sets and related benchmark resources.
[9] Society of Indian Automobile Manufacturers: Annual production and market reports, India.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY