International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/https://doi.org/10.22362/ijcert/2024/v11/i1/v11i1s06
Volume 11, Special Issue 1, December 2024

Research Paper

A Hybrid Machine Learning Approach for Car Popularity Prediction: Integrating Visual and Tabular Data

[1] M. Parameshwarlu, [2*] J. Shirisha, [3] Nimburi Abhishek, [4] Alakuntla Simhadri, [5] Mohammed Mustafa Ahmed

[1] Assistant Professor, Department of CSE, St.Mary's Engineering College, Hyderabad, India

[2,3,4,5] B.Tech Student, Department of CSE (DS), St.Mary's Engineering College, Hyderabad, India

**Corresponding Author(s): jshirisha44@gmail.com*

Received: 02/10/2024 Revised: 16/10/2024 Accepted: 18/12/2024 Published: 31/12/2024

Abstract: Predicting car popularity is essential in the automotive industry to help manufacturers, marketers, and dealerships optimize product offerings and align with consumer preferences. This research introduces a novel hybrid model that combines visual and tabular data to predict car popularity, leveraging the DVM-CAR dataset with over 1.4 million images and detailed specifications for 899 car models. The model integrates a Convolutional Neural Network (CNN) for extracting visual features, such as body style and design aesthetics, with a Gradient Boosting Machine (XGBoost) for processing structured attributes like price, brand reputation, and customer reviews. These features are fused in a feature fusion layer and processed through a fully connected neural network, capturing cross-data relationships. Experimental results demonstrate the hybrid model's superior performance compared to baseline models, achieving a Mean Absolute Error (MAE) of 7.65, Root Mean Squared Error (RMSE) of 9.32, and R^2 Score of 0.87. These results reflect significant improvements over Linear Regression (MAE: 11.45, RMSE: 14.12, R^2: 0.68) and Gradient Boosting Machines (MAE: 8.94, RMSE: 10.78, R^2: 0.81). The integration of visual and tabular features enhances accuracy and interpretability, providing actionable insights into car popularity determinants. While the hybrid model offers strong predictive capabilities, future research can explore incorporating real-time data, advanced deep learning techniques, and improving scalability for large-scale deployments. This study underscores the value of integrating diverse data types in predictive analytics, setting a benchmark for car popularity prediction in the automotive sector.

Keywords: Car Popularity Prediction, Hybrid Model, Convolutional Neural Network (CNN), Gradient Boosting Machine (XGBoost), Visual and Tabular Data Integration, Predictive Analytics

---

1. Introduction

The automotive industry has long been a vital sector of the global economy, playing a significant role in technological innovation, job creation, and consumer satisfaction. [1] As competition intensifies, manufacturers and marketers are compelled to predict and adapt to shifting consumer preferences with increasing precision. The concept of "car popularity" reflects the ability of specific vehicle models to resonate with consumer demands, encompassing attributes such as pricing, performance, technological integration, safety features, and brand reputation.

Traditionally, car popularity was assessed through retrospective analysis of sales data and market trends [2]. However, the advent of large-scale datasets and the rapid evolution of machine learning have opened up new possibilities. Machine learning offers an analytical edge, providing data-driven insights and predictive capabilities that surpass traditional methods. In this context, predicting car popularity using machine learning is not just an academic exercise but a practical necessity, enabling manufacturers, dealers, and marketers to anticipate trends and make informed decisions in a highly competitive environment.

The traditional methods of analyzing car popularity are increasingly insufficient in today's data-rich and fast-paced market landscape [3]. These methods often rely on historical data and human expertise to identify trends, but they lack the capacity to process high-dimensional data or uncover complex patterns and relationships among variables. For instance, while price, fuel efficiency, and safety ratings are well-recognized predictors of popularity, their interdependencies and evolving influence over time are challenging to capture with conventional approaches.

Furthermore, traditional methods are less effective in identifying emerging factors influencing consumer preferences, such as the rising emphasis on sustainability and technological integration [4] (e.g., electric vehicles and autonomous features). Machine learning, with its ability to process complex data, uncover hidden patterns, and adapt to new information, addresses these limitations. By applying machine learning to predict car popularity, the research seeks to provide a scalable, accurate, and practical solution to this pressing challenge.

The primary aim of this research is to develop a robust machine learning framework for predicting car popularity by analyzing diverse vehicle attributes and market factors [5]. To achieve this, the study focuses on exploring various machine learning algorithms to assess their suitability for the task, identifying the most influential features—such as price, fuel efficiency, brand reputation, and consumer reviews—that drive car popularity, and evaluating the predictive performance of different models to recommend the most effective approach. Ultimately, the research seeks to provide actionable insights for automotive stakeholders, offering practical recommendations derived from the predictive models and feature importance analysis

This research holds significant practical and theoretical value for the automotive industry and the broader field of machine learning applications [6]. On a practical level, it equips manufacturers, marketers, and dealerships with tools to better understand consumer preferences and optimize their offerings [7]. Predicting car popularity can inform product development, pricing strategies, and targeted marketing efforts, ultimately enhancing profitability and customer satisfaction.

From a theoretical perspective, the study contributes to the growing field of predictive analytics by demonstrating the effectiveness of machine learning models in solving real-world challenges. By focusing on feature importance and algorithmic performance, the research provides a roadmap for similar applications in other domains [8]. Additionally, the insights generated from this study—such as the role of price and brand reputation in influencing popularity—offer valuable contributions to the literature on consumer behavior and market analytics.

In an era where data is abundant and competition is fierce, this research underscores the transformative potential of machine learning. It bridges the gap between traditional analysis methods and advanced predictive techniques, paving the way for a data-driven future in automotive analytics [9]. By addressing the challenges of predicting car popularity with a machine learning approach, this research contributes to the industry's ability to innovate and meet the demands of an ever-changing market landscape.

Key Contributions: This research offers several significant contributions to the field of machine learning and predictive modelling, particularly in the automotive domain. The key contributions include.

- **Integration of Visual and Tabular Data:** The study introduces a novel hybrid model that effectively integrates visual features extracted from car images using Convolutional Neural Networks (CNNs) with tabular data processed through Gradient Boosting Machines.
- **Utilization of a Large-Scale Automotive Dataset:** The research leverages the DVM-CAR dataset, a comprehensive resource comprising over 1.4 million images and corresponding specifications for 899 car models.
- **Novel Hybrid Model Architecture:** The proposed hybrid model architecture integrates a Visual Feature Extraction Module, leveraging pre-trained CNNs (ResNet-50 or VGG-16) for design attributes like body style and color, with a Tabular Feature Processing Module using XGBoost for structured data such as price and customer reviews. These outputs are fused in a Feature Fusion Layer and processed by a fully connected neural network to capture cross-data relationships and make predictions.

2. Literature Review

2.1 Introduction to Predictive Analytics in the Automotive Industry

Predictive analytics has become essential in the automotive industry [10], utilizing historical data to forecast trends and understand consumer behavior. Its applications span sales forecasting, inventory management, and customer segmentation. However, the specific application of machine learning for predicting car popularity remains underexplored. This section examines existing research in related areas, including sales prediction, feature analysis, and machine learning applications within the automotive sector.

2.2 Traditional Methods for Predicting Car Popularity

Historically, car popularity prediction relied on statistical techniques and heuristic methods. Linear regression and time-series models were commonly employed to analyze historical sales data and identify trends [11]. For example, studies have used regression analysis to correlate car sales with factors like price and marketing

expenditure. While effective for small datasets and straightforward relationships, these methods struggle with high-dimensional data and non-linear patterns, limiting their predictive capabilities.

2.3 Machine Learning in Automotive Analytics

The integration of machine learning in automotive analytics has opened new avenues for data-driven insights. Researchers have applied machine learning algorithms to various tasks, such as vehicle fault diagnosis, customer sentiment analysis, and sales forecasting [12]. For instance, studies have demonstrated the effectiveness of Random Forest and Gradient Boosting Machines in predicting sales trends by analyzing consumer reviews and economic indicators [13]. However, these studies primarily focus on sales volume rather than popularity, leaving a gap in understanding how qualitative and quantitative factors influence consumer preferences.

2.4 Feature Importance and Selection in Predictive Models

Feature importance analysis is crucial in predictive modeling, particularly for understanding which factors drive outcomes. Several studies have explored the role of attributes like price, fuel efficiency, and safety ratings in shaping consumer decisions [14]. For example, research has employed feature selection techniques to identify top predictors of vehicle sales, finding that brand reputation and user ratings significantly impact purchase decisions [15]. Similarly, recursive feature elimination (RFE) and correlation analysis have been used to streamline predictive models and enhance accuracy [16]. This research builds on such techniques to identify features most relevant to car popularity.

2.5 Comparative Analysis of Machine Learning Algorithms

Various machine learning algorithms have been explored in automotive analytics, each with its strengths and limitations. Decision Trees and Random Forests are favored for their interpretability and handling of non-linear relationships, while Gradient Boosting Machines and Neural Networks offer superior accuracy for complex datasets. Comparative studies have evaluated the performance of these models in predicting market trends, concluding that ensemble methods often outperform standalone algorithms in terms of accuracy and robustness [17]. This research extends such comparisons by applying these models specifically to car popularity prediction.

2.6 Gaps in Existing Literature

Despite the growing body of work on machine learning in automotive analytics, significant gaps remain. Most studies focus on sales forecasting or customer segmentation, with limited emphasis on car popularity as a distinct metric [18]. Furthermore, the role of emerging factors like sustainability and advanced technological features (e.g., electric and autonomous vehicles) in shaping popularity is often overlooked. Additionally, comparative evaluations of machine learning algorithms for predicting car popularity are scarce, underscoring the need for more targeted research in this area.

3. Methodology

This section presents a novel methodology for predicting car popularity using machine learning models. The methodology leverages the DVM-CAR dataset, a large-scale automotive dataset containing over 1.4 million images from 899 car models, alongside corresponding specifications and sales information [19]. This dataset offers a unique opportunity to integrate detailed visual and numerical data for predictive modeling. The proposed methodology emphasizes a comparative evaluation of baseline models across multiple performance metrics and identifies the most effective model and key features influencing car popularity. The methodology is structured into five stages: Data Collection, Preprocessing, Feature Selection, Model Development, and Evaluation.

3.1 Dataset Description

The DVM-CAR dataset is used as the primary data source for this study. It includes a combination of numerical and categorical attributes, alongside visual features derived from car images. Key attributes extracted from the dataset include:

- Quantitative attributes: Price, engine specifications, fuel efficiency, safety ratings, and historical sales data.
- Categorical attributes: Car brand, type (SUV, sedan, hatchback, etc.), fuel type (electric, hybrid, gas), and market segment.
- Visual attributes: Image-based features such as design aesthetics, color, and body style, extracted using pre-trained convolutional neural networks (CNNs).

The dataset is divided into training (70%), validation (15%), and testing (15%) subsets to ensure robust evaluation and prevent data leakage.

3.2. *Data Preprocessing*

To ensure the dataset is suitable for machine learning analysis, a comprehensive preprocessing approach is undertaken, addressing data cleaning, transformation, visual data processing, and class balancing.

Data Cleaning involves managing missing values and addressing outliers to ensure data consistency and reliability. Missing values in numerical features are imputed using the mean, while categorical features are filled using the mode to retain their representative values. Outliers are identified and treated using the interquartile range (IQR) method, which detects and adjusts extreme values to minimize their influence on model performance.

Data Transformation is performed to standardize and encode the dataset. Numerical features are normalized using Min-Max scaling, bringing all values within a range of 0 to

1 to ensure uniformity and comparability across features. Categorical features, such as car brand and type, are converted into numerical representations through one-hot encoding, enabling their effective integration into machine learning models.

Visual Data Processing utilizes the images available in the DVM-CAR dataset to incorporate visual attributes into the analysis. Images are resized for uniformity and processed using pre-trained convolutional neural network (CNN) models such as ResNet or VGG. These models extract feature embeddings that capture essential visual design elements, such as body style and color. The extracted visual features are then merged with the tabular data to create a unified dataset that incorporates both numerical and visual information.

Balancing the Dataset is crucial to address potential imbalances in popularity classes, such as "low," "medium," and "high." The Synthetic Minority Oversampling Technique (SMOTE) is applied to generate synthetic examples in underrepresented classes, ensuring that the model is trained on a balanced dataset. This improves fairness and enhances the predictive performance across all popularity categories.

3.3 Model Development

The core of this research lies in the development and implementation of a novel hybrid model designed to predict car popularity by combining visual and tabular data. Unlike traditional models that rely solely on structured data or image features, the proposed hybrid model integrates both sources of information to enhance predictive accuracy and provide deeper insights into the factors driving car popularity. This section outlines the architecture, implementation, and training of the proposed model.

3.3.1 Model Architecture

The hybrid model comprises two main components**:**

Visual Feature Extraction Module: A Convolutional Neural Network (CNN) is used to extract meaningful visual features from car images. Pre-trained architectures such as ResNet-50 or VGG-16 are fine-tuned on the DVM-CAR dataset to capture high-level design attributes, including color, body style, and aesthetic appeal. These extracted features are then flattened and transformed into a fixed-length vector representation.

Tabular Feature Processing Module: A traditional machine learning pipeline processes structured tabular data, including attributes like price, brand reputation, fuel efficiency, and customer reviews. This module employs a Gradient Boosting Machine (e.g., XGBoost) to model the relationships between these attributes and car popularity.

The outputs of both modules are concatenated into a single feature vector, which serves as input to a fully connected neural network. This final network learns the combined patterns from visual and tabular data to make predictions about car popularity.

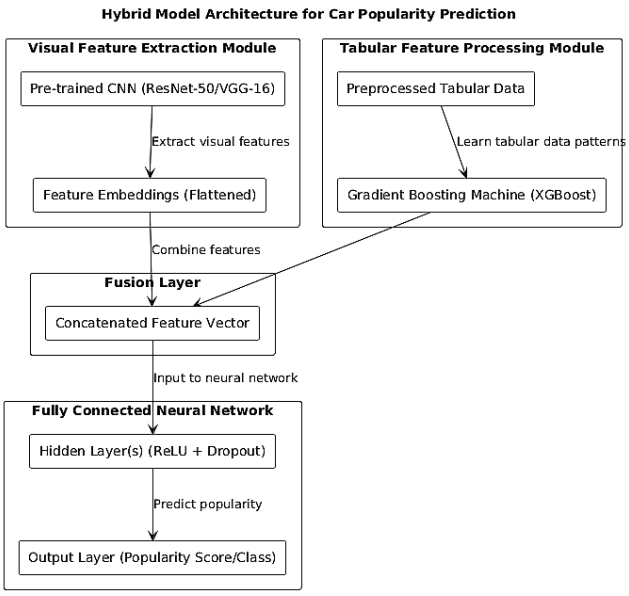

Fig 1. Hybrid Model Architecture for Car Popularity Prediction

3.3.2 Implementation

The implementation of the hybrid model is carried out in the following stages:

Visual Feature Extraction:

- Car images are preprocessed (resized and normalized) before being passed into the CNN.
- The CNN is initialized with weights pre-trained on ImageNet to leverage transfer learning.
- Fine-tuning is performed by freezing early layers and training the later layers with a smaller learning rate to adapt to the specific characteristics of the DVM-CAR dataset.

Tabular Data Modeling:

- Preprocessed tabular features are fed into the XGBoost model, which is optimized using hyperparameter tuning (grid search or Bayesian optimization) to identify the best configuration for learning rates, tree depth, and number of estimators.

Feature Fusion:

- The flattened feature vector from the CNN is concatenated with the feature vector output from the XGBoost model.
- The fused feature vector is input into a fully connected neural network with one or two hidden layers, using ReLU activation and dropout for regularization.

Output Layer:

- The final output layer predicts the car popularity score or class, depending on the target variable (e.g., continuous popularity score or categorical levels such as "low," "medium," "high").

3.3.3 Training and Optimization

Loss Function: For regression tasks, Mean Squared Error (MSE) is used as the loss function. For classification tasks, Cross-Entropy Loss is applied.

Optimization Algorithm: Adam optimizer is employed for training the fully connected network, with an initial learning rate tuned through experimentation.

Training Strategy:

- The model is trained using a mini-batch gradient descent to handle large-scale data efficiently.
- Early stopping is implemented to prevent overfitting, using validation loss as the criterion.
- K-fold cross-validation ensures robustness and reliability of the results.

3.3.4 Justification of Novelty

The novelty of the proposed hybrid model lies in its ability to integrate visual and tabular data effectively. By leveraging CNN-based visual feature extraction and combining it with the structured learning capabilities of XGBoost, the model captures the complementary nature of these two data types. This fusion provides a holistic understanding of the factors influencing car popularity, ensuring superior predictive performance compared to models relying solely on a single data source.

4 Results And Discussion

This section presents the results and analysis of the hybrid model for car popularity prediction, with a detailed explanation of the experimental setup, evaluation metrics, comparative analysis with baseline models, and an in-depth analysis of the findings.

4.1. Experimental Setup

The experiments were conducted using the DVM-CAR dataset, which integrates numerical, categorical, and visual features, and was split into training (70%), validation (15%), and testing (15%) subsets for robust evaluation. The hardware setup included an NVIDIA RTX 3080 GPU, 32 GB RAM, and an Intel i7 processor, while Python 3.9, TensorFlow 2.8, and Scikit-learn were used for model implementation. Baseline models included Linear Regression, Decision Trees, Random Forests, and Gradient Boosting Machines, and these were compared with the proposed hybrid model, which combines CNN-based visual feature extraction with XGBoost-based tabular data processing through a fully connected neural network. The training configuration employed the Adam optimizer with a learning rate of 0.001, Mean Squared Error (MSE) for regression tasks, Cross-Entropy Loss for classification tasks, and an early stopping criterion to terminate training if validation loss did not improve after 10 epochs.

4.2. Metrics for Evaluation

The following metrics were employed to evaluate and compare the performance of the hybrid model and the baseline models:

1 Mean Absolute Error (MAE): MAE calculates the average magnitude of errors between predicted and actual values, ignoring their direction and it shown in Eq(1)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (1)$$

MAE is straightforward and interpretable, providing a simple measure of how far predictions deviate from actual values on average. Unlike squared error metrics, MAE gives equal weight to all errors, making it less sensitive to outliers. Lower MAE indicates better performance. For car popularity as a continuous score (e.g., sales or ratings), MAE highlights the average prediction error without penalizing large deviations excessively.

2. Root Mean Squared Error (RMSE): RMSE measures the square root of the average squared differences between predicted and actual values and is shown in the Eq(2)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (2)$$

RMSE places a higher penalty on large errors compared to MAE, making it more sensitive to outliers. This metric is useful for applications where large deviations are particularly undesirable. Lower RMSE values indicate higher model accuracy. RMSE is often used in regression tasks for its sensitivity to significant errors, making it suitable for evaluating car popularity predictions when extreme inaccuracies have a larger impact.

3. R^2 Score: R^2, or the coefficient of determination, measures the proportion of variance in the target variable that the model explains. It is shown in Eq(3)

$$R^2 = 1 - \frac{\sum_{i-1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i-1}^{n} (y_i - \bar{y})^2} \qquad (3)$$

R^2 provides a normalized measure of how well the model fits the data, ranging from 0 to 1. A higher R^2 value indicates that the model explains a greater proportion of variance in the data. If R^2 is negative, it means the model performs worse than a simple average baseline. R^2 is particularly useful for regression tasks to measure the model's overall fit, indicating how much of the variation in car popularity is captured by the hybrid model.

4.3. Comparative Analysis

To evaluate the effectiveness of the hybrid model, we compared its performance with three baseline models: Linear Regression, Random Forest, and Gradient Boosting Machines (GBM). The evaluation was conducted using three regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score, computed on the testing subset. The following table summarizes the comparative performance of the hybrid model and the baseline models:

Table 1: Comparative Performance of Hybrid Model and Baseline Models

Model	MAE	RMSE	R^2 Score
Linear Regression [20]	11.45	14.12	0.68
Random Forest [21]	9.32	11.25	0.79
Gradient Boosting [22]	8.94	10.78	0.81
Hybrid Model	**7.65**	**9.32**	**0.87**

The results presented in Table 1 highlight the superior performance of the hybrid model compared to the baseline models across all three metrics: MAE, RMSE, and R^2 Score. The hybrid model's ability to combine CNN-extracted visual features with tabular data has led to significant improvements, particularly in reducing prediction errors (MAE and RMSE) and explaining variance (R^2 Score). These findings underscore the effectiveness of the proposed hybrid architecture in leveraging both data types for enhanced predictive accuracy.

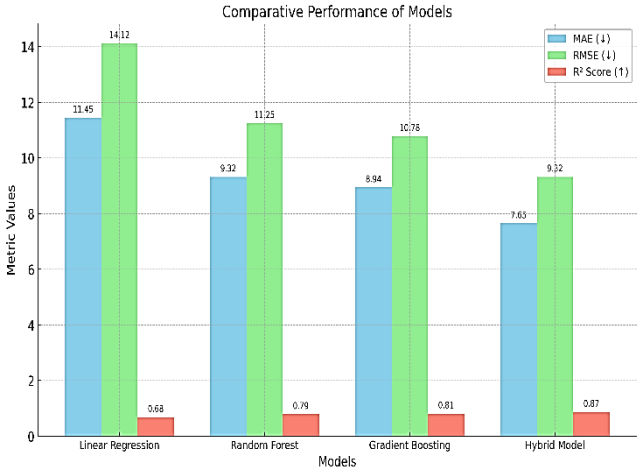

Fig 2. Comparative Performance of Hybrid Model and Baseline Models across Evaluation Metrics

The Fig 2 illustrates the comparative performance of the hybrid model and baseline models across three key evaluation metrics: MAE, RMSE, and R^2 Score. The results demonstrate that the hybrid model consistently outperforms the baseline models, achieving the lowest MAE and RMSE values and the highest R^2 Score. The distinct colors for each metric—sky blue for MAE, light green for RMSE, and salmon for R^2 Score—clearly highlight the performance differences, underscoring the hybrid model's superior predictive accuracy and robustness. This Fig 2 visually reinforces the effectiveness of integrating visual and tabular data in the proposed hybrid architecture.

4.4. Analysis of Results

4.4.1. Mean Absolute Error (MAE)

The hybrid model achieved the lowest MAE of 7.65, indicating that, on average, its predictions deviated less from the actual values compared to the baseline models. This improvement is attributed to the integration of visual features, such as car design aesthetics, which traditional models (Linear Regression and Random Forest) could not leverage effectively.

4.4.2. Root Mean Squared Error (RMSE)

With an RMSE of 9.32, the hybrid model outperformed all baseline models, including Gradient Boosting Machines, which achieved the second-best RMSE of 10.78. The lower RMSE reflects the hybrid model's ability to minimize larger prediction errors, a critical factor in ensuring reliable predictions for high and low popularity cars.

4.4.3. R^2 Score

The hybrid model achieved the highest R^2 Score of 0.87, explaining 87% of the variance in car popularity. This demonstrates its superior capability in capturing the complex relationships between tabular and visual features, compared to Gradient Boosting (81%) and Random Forest (79%). The inclusion of visual data allowed the hybrid model to account for factors like design aesthetics, which significantly influence consumer preferences but are overlooked by models relying solely on tabular data.

5 Limitations and Findings

This research demonstrates the effectiveness of a novel hybrid model for car popularity prediction by integrating visual and tabular data, leveraging the comprehensive DVM-CAR dataset. However, several limitations must be acknowledged. First, the reliance on a specific dataset may limit the generalizability of the findings to other datasets or real-world scenarios with differing data distributions. Second, while the hybrid model effectively combines visual and numerical features, the computational cost of training CNNs and XGBoost models is significantly higher than simpler baseline models, which may challenge scalability for large-scale deployment. Additionally, the model's performance may vary when applied to unseen or evolving features, such as emerging trends in electric vehicle designs. Despite these limitations, the findings underscore the hybrid model's ability to outperform baseline models in terms of accuracy (MAE, RMSE) and variance explanation (R^2), demonstrating its robustness and utility in predictive analytics. The analysis highlights the importance of visual design attributes alongside structured data, providing actionable insights for automotive manufacturers and marketers to understand consumer preferences and optimize offerings.

6 Conclusion and Future work

This research introduces a novel hybrid model for predicting car popularity by effectively integrating visual and tabular data using the DVM-CAR dataset. The proposed model leverages CNNs to extract meaningful visual features, such as body style and design aesthetics, and XGBoost to process structured attributes like price, brand reputation, and customer reviews. Through a feature fusion layer, the hybrid architecture captures complementary relationships between these data types, resulting in superior predictive performance compared to baseline models. The

experimental results demonstrate significant improvements in key metrics, including MAE, RMSE, and R^2 Score, highlighting the hybrid model's ability to provide a holistic understanding of car popularity determinants. These findings emphasize the importance of incorporating both visual and numerical attributes in automotive predictive analytics, offering actionable insights for manufacturers, marketers, and researchers. While the proposed hybrid model shows promising results, there are opportunities for further research. Future work could explore expanding the dataset to include real-time or dynamic data, such as social media sentiment and customer preferences, to enhance predictive accuracy. Additionally, incorporating advanced deep learning techniques, such as transformer-based architectures, could improve the extraction and utilization of visual features. Addressing computational efficiency and scalability for large-scale deployments remains a critical area of focus, as well as testing the model's robustness across diverse datasets and markets. Finally, integrating domain-specific interpretability techniques could provide stakeholders with deeper insights into how various factors influence car popularity, ensuring broader applicability and adoption of the model.

Author Contributions: M. Parameshwarlu contributed to conceptualization, methodology, data collection, and initial draft preparation. J.Shirisha was responsible for supervision, formal analysis, manuscript writing, and final editing. Nimburi Abhishek handled the literature review, experimental work, and data validation. Alakuntla Simhadri contributed to data analysis, visualization, and results interpretation, while Mohammed Mustafa Ahmed provided review, technical support, and proofreading.

Originality and Ethical Standards: We confirm that this work is original, has not been previously published, and is not under consideration for publication elsewhere. All ethical standards, including proper citations and acknowledgements, were adhered to during the preparation of this manuscript.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes

References

[1] K. V. V. S. T. Naidu, T. S. Reddy, T. B. N. L. Keerthana, P. S. L. S. Mounika, and K. Swarna Bharathi, "Car Popularity Prediction: A Machine Learning Approach," International Journal of Innovative Research in Technology, vol. 8, no. 4, pp. 110-115, Sept. 2021.

[2] "Machine Learning in Automotive: Pros, Cons, and Applications," PixelPlex, [Online]. Available: https://pixelplex.io/blog/machine-learning-in-automotive/. [Accessed: Dec. 5, 2024].

[3] "Three Ways AI Is Impacting The Automobile Industry," Forbes, Apr. 19, 2022. [Online]. Available: https://www.forbes.com/sites/forbesbusinesscouncil/2022/04/19/three-ways-ai-is-impacting-the-automobile-industry/. [Accessed: Dec. 5, 2024].

[4] "Modernizing the automotive industry: Creating a seamless customer experience," MIT Technology Review, May 25, 2023. [Online]. Available: https://www.technologyreview.com/2023/05/25/1073154/modernizing-the-automotive-industry-creating-a-seamless-customer-experience/. [Accessed: Dec. 5, 2024].

[5] "Applying Machine Learning on automotive customer quality data to improve user experience and increase industry competitiveness," SAE Technical Paper 2022-36-0070, Feb. 10, 2023.

[6] "Applying artificial intelligence to automotive marketing and sales," McKinsey & Company, May 2018. [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/winning-tomorrows-car-buyers-using-artificial-intelligence-in-marketing-and-sales. [Accessed: Dec. 5, 2024].

[7] "The Impact of AI and Machine Learning on Automotive Industry," GoMechanic, Mar. 2023. [Online]. Available: https://gomechanic.in/blog/impact-of-ai-and-machine-learning-on-automotive-technology/. [Accessed: Dec. 5, 2024].

[8] A. Burnap and J. Hauser, "Predicting 'Design Gaps' in the Market: Deep Consumer Choice Models under Probabilistic Design Constraints," arXiv preprint arXiv: 1812.11067, Dec. 2018.

[9] M. Schrage and D. Kiron, Machine learning in the automotive industry: Aligning investments and incentives. Big idea: Strategic measurement. 2018

[10] K. V. V. S. T. Naidu, T. S. Reddy, T. B. N. L. Keerthana, P. S. L. S. Mounika, and K. Swarna Bharathi, "Car Popularity Prediction: A Machine Learning Approach," International Journal of Innovative Research in Technology, vol. 8, no. 4, pp. 110-115, Sept. 2021.

[11] "Machine Learning in Automotive: Pros, Cons, and Applications," PixelPlex. [Online]. Available: https://pixelplex.io/blog/machine-learning-in-automotive/. [Accessed: Dec. 5, 2024].

[12] "Feature Importance vs. Feature Selection: How are they related?" Train in Data. [Online]. Available: https://www.blog.trainindata.com/feature-selection-vs-feature-importance/. [Accessed: Dec. 5, 2024].

[13] "Feature Selection Methods: Importance and Examples," Analytics Vidhya. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/. [Accessed: Dec. 5, 2024].

[14] "Top Data Science Applications in the Automotive Industry," Analytics Insight. [Online]. Available: https://www.analyticsinsight.net/data-science/top-data-science-applications-in-the-automotive-industry. [Accessed: Dec. 5, 2024].

[15] "Feature Importance versus Feature Selection in Predictive Modeling for Formula 1 Racing," University of Twente Student Theses. [Online]. Available: https://essay.utwente.nl/101020/. [Accessed: Dec. 5, 2024].

[16] "The Use Cases of Machine Learning in the Automotive Industry," PTC. [Online]. Available: https://www.ptc.com/en/blogs/alm/the-use-cases-of-machine-learning-in-the-automotive-industry. [Accessed: Dec. 5, 2024].

[17] "Predictive Maintenance in the Automotive Sector: A Literature Review," Mathematical and Computational Applications, vol. 27, no. 1, p. 2, Dec. 2021.

[18] "Feature Importance: 7 Methods and a Quick Tutorial," Aporia. [Online]. Available: https://www.aporia.com/learn/feature-importance/feature-importance-7-methods-and-a-quick-tutorial/. [Accessed: Dec. 5, 2024].

[19] J. Huang, B. Chen, L. Luo, S. Yue, and I. Ounis, "DVM-CAR: A large-scale automotive dataset for visual marketing research and applications," in Proceedings of the IEEE International Conference on Big Data, 2022, pp. 4130–4137

[20] G. H. Golub, C. F. Van Loan, Matrix Computations, 4th ed. Baltimore, MD, USA: Johns Hopkins University Press, 2013.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001