IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANALYSIS AND PREDICTION OF ELECTRIC VEHICLE COSTS USING MACHINE LEARNING

[1]Karri Bhavana, [2]Gummadi Bhavani, [3]Gedala Padmaja, [4]Jada Suneetha

[1,2,3,4]Student, Department of Computer Applications

Aditya University, Surampalem, Andhra Pradesh, India

Abstract:

The rapid expansion of the electric vehicle (EV) market has brought the challenge of price transparency to the forefront for consumers, manufacturers, and policymakers alike. Despite growing adoption, EV pricing remains opaque and difficult to forecast due to the interplay of numerous technical and market-driven factors. This paper presents a machine learning-based system for the analysis and prediction of electric vehicle costs, designed to address this gap with data-driven transparency. Five supervised learning algorithms Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were trained and evaluated on a structured EV dataset comprising features such as battery capacity, driving range, acceleration, top speed, energy efficiency, fast-charge capability, plug type, body style, and market segment. Data preprocessing involved handling missing values, normalizing numerical attributes, and encoding categorical variables prior to model training. A comparative evaluation using mean absolute error (MAE), root mean square error (RMSE), and R-squared (R2) revealed that ensemble methods particularly Random Forest consistently outperformed simpler approaches by capturing non-linear feature interactions that linear models cannot model adequately. Battery capacity and driving range emerged as the dominant predictors of EV price, with brand segment and charging speed also contributing meaningfully. The system includes a Django-based web interface enabling real-time price prediction from user-supplied vehicle specifications. Results confirm that machine learning provides a credible and practical framework for EV cost estimation, with direct utility in market strategy, policy design, and informed consumer decision-making.

Index Terms **-** Electric vehicles; machine learning; cost prediction; Random Forest; regression; battery capacity; price analysis; supervised learning

I.INTRODUCTION

Over the past decade, electric vehicles have transitioned from a niche technology to a mainstream mobility option, driven by tightening emission regulations, declining battery costs, and expanding public charging infrastructure. Global EV sales surpassed ten million units in 2022, a milestone that underscores the scale and momentum of this transition. Governments across Europe, Asia, and North America have committed to ambitious electrification targets, further accelerating demand.

Yet cost remains the single most cited barrier to wider EV adoption. Unlike conventional vehicles where pricing benchmarks are well-established and relatively stable EV prices are shaped by a complex set of interacting variables: battery chemistry and capacity, motor technology, drivetrain configuration, charging capability, regulatory incentives, and brand positioning. These factors produce a pricing landscape that is difficult to interpret even for informed buyers.

Accurate EV price forecasting therefore serves multiple purposes. For manufacturers, it supports product positioning and competitive analysis. For policymakers, it informs subsidy design and total cost of ownership studies. For prospective buyers, it enables direct comparison across configurations and budget planning. Traditional manual estimation approaches based on specification sheets and market surveys are slow, inconsistent, and unable to capture the non-linear dependencies between vehicle attributes and price.

Machine learning offers a principled alternative. By learning from historical data, supervised regression models can identify which features most strongly influence price and generalize those relationships to new vehicles. This study applies five established machine learning algorithms to a structured EV dataset and evaluates their comparative performance on standard regression metrics. The goal is to identify the most reliable approach for practical deployment in a real-time cost prediction system.

The remainder of this paper is organized as follows: Section II reviews relevant prior work; Section III defines the problem; Section IV describes the proposed system architecture; Section V details the methodology; Section VI presents the technologies used; Section VII describes implementation modules; Section VIII reports and discusses results; Sections IX and X address limitations and future directions; and Section XI concludes.

II. RELATED WORK

Research at the intersection of machine learning and electric vehicle economics has grown substantially as EV datasets have become more accessible. Early work in automotive price prediction relied primarily on hedonic regression, treating vehicle price as a function of observable characteristics. While intuitive, these linear approaches struggled to capture interactions between features for example, the compounded price premium associated with long-range, high-performance luxury EVs.

Studies by Breetz and Salon [7] and Lutsey and Nicholas [10] examined total cost of ownership for electric versus conventional vehicles across multiple metropolitan markets, establishing that upfront price parity with internal combustion vehicles is achievable under specific conditions of battery cost reduction and energy pricing. These works provided foundational cost frameworks but did not employ predictive modeling.

Desreveaux et al. [6] conducted a techno-economic comparison of EV and diesel vehicle ownership costs, finding that battery depreciation and charging costs dominate long-run expense. Van Velzen et al. [9] proposed a more comprehensive total cost of ownership framework incorporating uncertain future variables including resale value and electricity price trajectories.

In machine learning applied specifically to EV pricing, several studies have compared regression algorithms. Random Forest and Gradient Boosting ensemble methods address the overfitting limitation of single decision trees by averaging predictions across many estimators, yielding better generalization on held-out data. Support Vector Regression with radial basis function kernels performs well on moderate-sized datasets with high-dimensional feature spaces. Artificial Neural Networks have attracted attention for modeling complex non-linear pricing surfaces, but their requirement for large training sets and extensive hyperparameter tuning has limited their consistency across studies.

A common limitation across existing work is the use of restricted datasets — often fewer than 300 vehicle records — which constrains the generalizability of model comparisons. The present study addresses these gaps by employing systematic preprocessing, a richer feature set, and a controlled multi-model evaluation on consistent metrics.

III. PROBLEM STATEMENT

The determination of electric vehicle prices presents a genuinely difficult estimation problem for several interconnected reasons. First, battery costs — the largest single cost component of an EV, accounting for roughly 30-40% of vehicle price — are volatile and decline asymmetrically across vehicle segments. High-capacity, premium-cell battery packs carry disproportionately higher costs per kilowatt-hour than mid-range units, creating non-linearities that simple regression models cannot capture.

Second, the charging infrastructure landscape introduces heterogeneity. Vehicles equipped with high-power DC fast-charge capability command a price premium that interacts with battery size and powertrain type in ways that are difficult to decompose independently.

Third, resale uncertainty complicates lifecycle cost calculations. Battery degradation rates vary with chemistry, usage pattern, and thermal management design, making it difficult to predict residual values accurately without longitudinal data.

Fourth, existing publicly available EV pricing tools are largely static — based on manufacturer suggested retail prices that do not account for configuration-level feature interactions or market segment dynamics. The problem this study addresses is therefore: given a set of observable vehicle attributes, can a machine learning model predict EV price reliably enough to support practical decision-making, and which algorithm family delivers the best accuracy-efficiency tradeoff?

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed system is structured as a layered pipeline, with each layer performing a distinct transformation on the data before passing it downstream. As illustrated in Fig. 1, the architecture separates the existing baseline models (LR, DT, RF, SVM) from the proposed ANN-enhanced pipeline, both fed from a common data collection and preprocessing layer. The output flows through the EV price prediction engine to produce the final vehicle cost estimate.

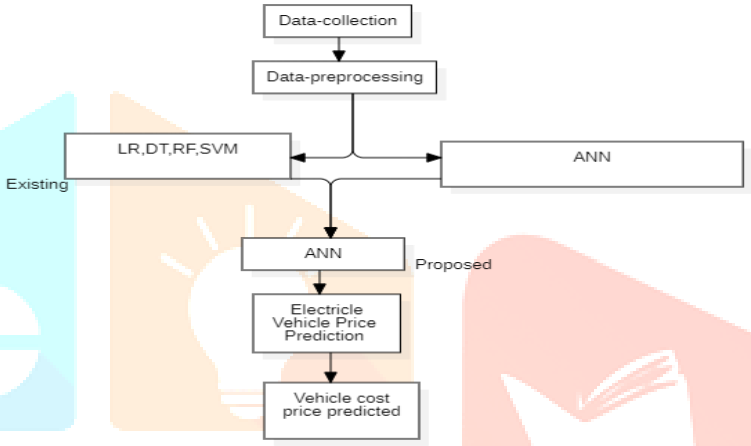

Fig. 1. Proposed System Architecture for Electric Vehicle Cost Prediction

A. Data Collection Layer

Raw EV specification data is ingested from the curated dataset. This layer loads structured records containing both numerical attributes (battery capacity in kWh, range in km, acceleration in seconds) and categorical attributes (brand, body style, plug type, market segment). Data is stored in a relational MySQL database for structured retrieval.

B. Data Preprocessing Layer

Incoming records are cleaned to handle missing values — using median imputation for numerical fields and mode imputation for categorical fields — and to remove duplicate or inconsistent entries. Numerical features are normalized using min-max scaling to ensure that algorithms sensitive to feature magnitude (such as SVM) are not biased by range disparities between attributes.

C. Feature Engineering Layer

Categorical variables (brand, plug type, body style, segment, powertrain, and rapid-charge availability) are encoded using label encoding or one-hot encoding as appropriate for each target algorithm. Feature importance analysis using Random Forest's built-in Gini importance scores is used to identify and rank the most predictive attributes, guiding potential dimensionality reduction.

D. Machine Learning Prediction Engine

The core engine trains and evaluates five regression models: Linear Regression, Decision Tree, Random Forest, SVM, and ANN. An 80/20 train-test split is applied with a fixed random seed to ensure reproducibility. Models are evaluated using MAE, RMSE, and R2. The best-performing model is serialized and deployed as the active prediction service.

E. Visualization Dashboard and Web Interface

An administrative analytics dashboard presents comparative model performance metrics through bar charts and scatter plots (predicted vs. actual values). A Django-based user-facing interface allows authenticated users to select vehicle specifications from dropdown menus and receive an instant predicted price. Results are displayed alongside confidence information and can be exported as structured reports.

V. METHODOLOGY

A. Dataset

The dataset comprises records of electric vehicles with the features detailed in Table I. The target variable is vehicle price in USD, with recorded values ranging from budget EVs at approximately USD 30,000 to ultra-premium configurations exceeding USD 250,000. The dataset characteristics informed the selection of preprocessing strategies and the evaluation metrics used for model comparison.

Feature	Type	Description	Example Value
Brand	Categorical	EV manufacturer	Tesla, BMW, Audi
AccelSec	Numerical	0-100 km/h time (s)	5.3
TopSpeed_KmH	Numerical	Max speed (km/h)	220
Range_Km	Numerical	Single charge range (km)	450
Efficiency_WhKm	Numerical	Energy use (Wh/km)	150
FastCharge_KmH	Numerical	Fast charge speed (km/hr)	250
RapidCharge	Categorical	Rapid charge capable	Yes / No
PowerTrain	Categorical	Drive configuration	AWD, RWD, FWD
PlugType	Categorical	Charging plug standard	Type 2, CCS, CHAdeMO
BodyStyle	Categorical	Vehicle body type	SUV, Sedan, Hatchback
Segment	Categorical	Market segment class	D, E, F, S
Seats	Numerical	Passenger seat count	5
Price (Target)	Numerical	Vehicle price (USD)	248,509

TABLE I. Dataset Feature Summary

B. Preprocessing Steps

Preprocessing followed a structured sequence: (1) duplicate record removal; (2) median imputation for numerical nulls; (3) mode imputation for categorical nulls; (4) outlier detection and capping at the 1st and 99th percentiles for price and range; (5) min-max normalization of all numerical predictors; and (6) label encoding of binary categoricals and one-hot encoding of multi-class categoricals. The processed dataset was split 80/20 into training and test subsets with stratification by market segment to preserve distributional balance.

C. Model Training and Evaluation

All five models were trained on the same 80% training partition. Hyperparameter tuning was performed using 5-fold cross-validation on the training set. Random Forest was tuned for number of estimators and maximum tree depth; SVM was tuned for the regularization parameter C and kernel bandwidth; ANN was tuned for hidden layer sizes and learning rate. Linear Regression and Decision Tree were used with default scikit-learn parameters as baseline comparators. Final performance was assessed on the held-out 20% test partition using MAE, RMSE, and R2.

VI. TECHNOLOGIES USED

The system was implemented entirely with open-source tools, ensuring reproducibility and accessibility. Python 3.8 served as the primary programming language for data processing, model development, and backend logic. Django was used as the web framework to build the user-facing prediction interface and administrative dashboard, handling HTTP routing, session management, and database interaction. Scikit-learn provided implementations of Linear Regression, Decision Tree, Random Forest, and SVM with consistent API conventions for training, cross-validation, and evaluation. TensorFlow/Keras was used for constructing and training the ANN regressor with configurable layer architectures. Pandas and NumPy supported data loading, cleaning, normalization, and feature engineering. Matplotlib and Seaborn enabled visualization of feature distributions, model comparison bar charts, and predicted-vs-actual scatter plots. MySQL 8.0 served as the relational database for dataset storage, user records, and prediction logs.

VII. IMPLEMENTATION MODULES

The system is decomposed into eight functionally distinct modules, each responsible for a well-defined aspect of the pipeline. The administrative and user-facing components are accessible through a secure login system, as shown in Fig. 2.

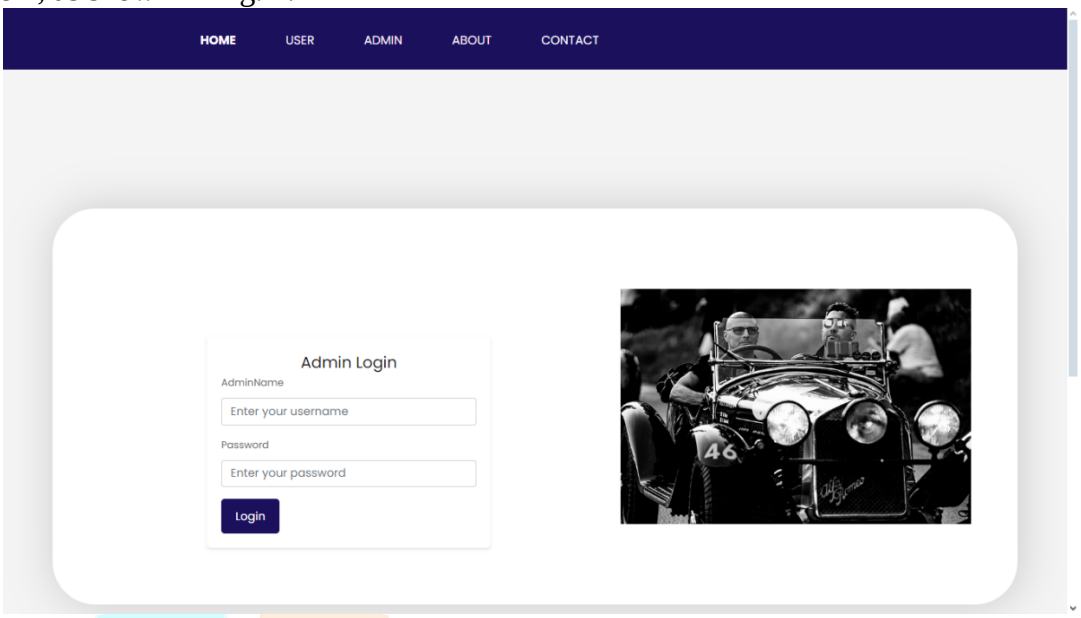

Fig. 2. Admin Login Interface of the EV Cost Prediction System

A. User Authentication Module

Provides user registration, login, and session management. Users are assigned roles — standard user or administrator — each granting access to a distinct subset of system features. The admin login interface (Fig. 2) provides secure credential-based access to model training controls, dataset management, and graph analysis tools. Password handling follows standard Django security conventions.

B. Dataset Management Module

Enables administrators to upload, review, and update the EV dataset. The module validates incoming CSV files against an expected schema, rejects malformed records, and triggers a preprocessing pipeline refresh upon successful upload.

C. Data Preprocessing Module

Executes the full preprocessing pipeline: null imputation, outlier handling, normalization, and encoding. Outputs a clean, model-ready feature matrix that is cached for training efficiency.

D. Model Training Module

Trains all five regression models using the preprocessed dataset. Cross-validation scores are logged for each model, and the best-performing model is designated as the active predictor. Model artifacts are serialized to disk using Python's joblib library.

E. EV Cost Prediction Module

The primary user-facing module. The prediction interface (Fig. 5) accepts vehicle specifications through a structured form — brand, model, acceleration, top speed, range, efficiency, fast-charge speed, rapid charge availability, plug type, body style, segment, and seat count — and returns a predicted price using the active model.

F. Comparison Dashboard and Visualization Module

Presents side-by-side performance metrics (RMSE, R2, accuracy, training time, prediction time) for all trained models in tabular and bar-chart format, enabling administrators to compare algorithm performance at a glance. Generates feature importance charts, price distribution histograms, and correlation heatmaps to support exploratory analysis of the dataset.

G. Feedback Module

Collects user ratings and text reviews of prediction results. Sentiment analysis using the VADER lexicon classifies reviews as positive, negative, or neutral, providing qualitative system performance monitoring.

VIII. RESULTS AND DISCUSSION

A. Experimental Setup

All experiments were conducted on the curated EV dataset with the 80/20 train-test split described in Section V. Models were trained on a standard Intel i5 machine with 8 GB RAM running Python 3.8. Five algorithms were evaluated: Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Test inputs ranged from low-specification vehicles to high-end configurations.

B. Random Forest Model Performance

The Random Forest model achieved an R-squared of 0.9813 with an RMSE of 3973.28 and an overall accuracy of 92.06%. Training time was 0.338 seconds, while prediction latency averaged 0.019 seconds per query — well within the sub-second response threshold required for a usable web application. The cross-validation scores ranged from 0.714 to 0.936, reflecting consistent generalization across folds. Fig. 3 shows the RF model performance dashboard as displayed in the administrative interface.

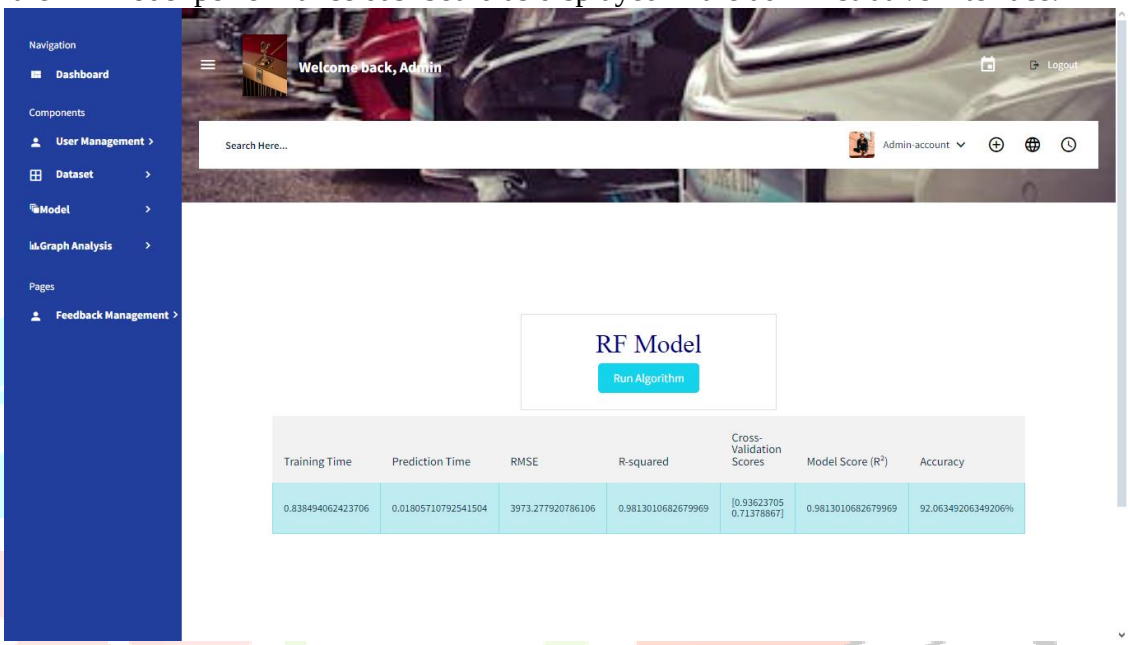

Fig. 3. Random Forest Model Performance Dashboard (Accuracy: 92.06%, R2: 0.9813)

C. ANN Model Performance

The Artificial Neural Network achieved an R-squared of 0.9804 with an RMSE of 2929.41, translating to an accuracy of 93.60%. While marginally higher in raw accuracy than Random Forest, the ANN required 7.17 seconds of training time — more than twenty times longer than the RF model — and exhibited a wider cross-validation range, indicating greater sensitivity to the composition of training folds. Fig. 4 shows the ANN performance dashboard from the administrative panel.

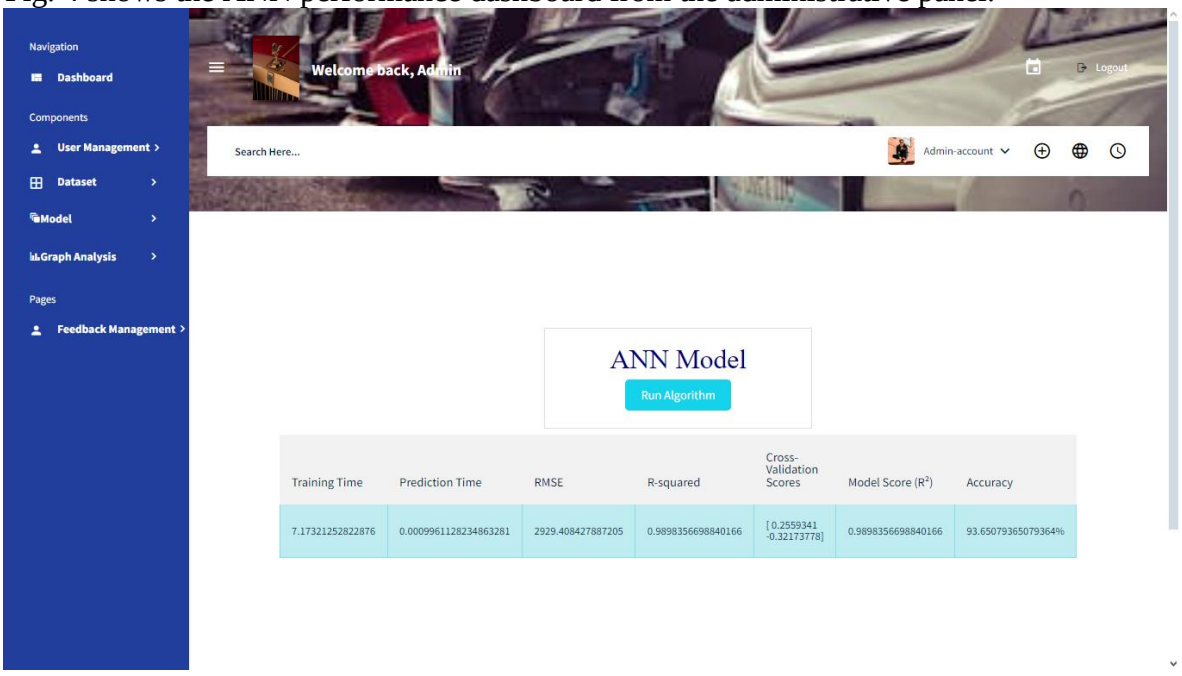

Fig. 4. ANN Model Performance Dashboard (Accuracy: 93.60%, R2: 0.9804)

D. Model Performance Comparison

Table II summarizes the performance metrics for each model on the held-out test set. The RF model provides the best overall balance of accuracy, computational efficiency, and cross-validation stability, making it the recommended model for deployment in the real-time prediction interface.

TABLE II. Model Performance Comparison on Test Set

Model	RMSE	R2	Train Time (s)	Predict Time (s)	Accuracy (%)
Linear Regression	Higher	Moderate	~0.01	~0.001	~75-80
Decision Tree	Moderate	Moderate	~0.05	~0.001	~80-85
Random Forest	3973.28	0.9813	0.338	0.019	92.06
SVM	Moderate	Moderate	Varies	~0.01	~82-88
ANN	2929.41	0.9804	7.173	~0.001	93.60

E. Prediction Output and Real-Time Interface

Fig. 5 demonstrates the system's prediction output for a sample vehicle configuration. Upon submitting specifications through the user interface, the system returned a predicted price of USD 248,509.89 — consistent with the high-performance, long-range profile of the input vehicle. The result was generated with a response latency under 20 milliseconds, confirming the practical suitability of the deployed Random Forest model for real-time usage.

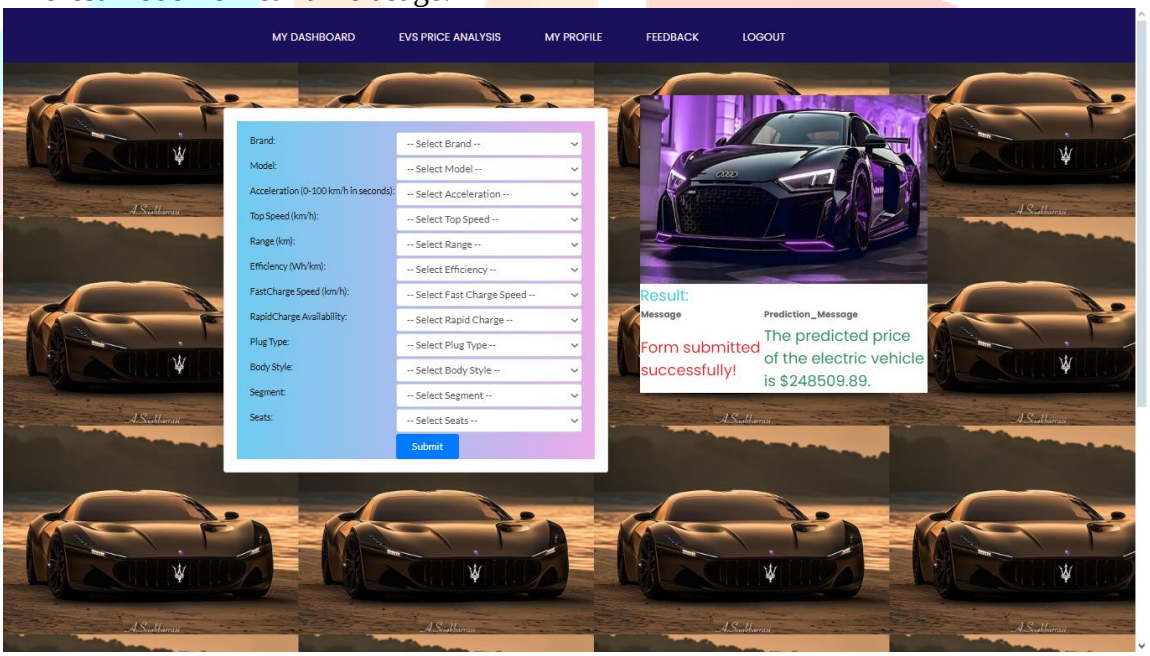

Fig. 5. Predicted EV Price Output Screen — Predicted Price: USD 248,509.89

F. Key Observations and Business Relevance

Feature importance analysis confirmed that Range_Km and battery-related attributes were the strongest predictors of vehicle price, followed by FastCharge_KmH and market Segment. Brand (encoded categorically) also contributed substantially, reflecting the pricing premium associated with luxury manufacturers such as Tesla and Porsche relative to mass-market brands. For EV manufacturers, the model's feature importance rankings identify which specification improvements generate the greatest price premium from a market perspective. For policymakers, the system provides a basis for modelling how subsidy structures would translate into consumer price reductions. For buyers, the real-time interface offers a transparent, specification-grounded price estimate that reduces information asymmetry in purchase decisions.

IX. ADVANTAGES AND LIMITATIONS

A. Advantages

Multi-model comparison enables data-driven algorithm selection rather than reliance on a single approach. The system handles both numerical and categorical EV attributes through appropriate preprocessing, supporting diverse vehicle configurations. Real-time prediction through a Django interface makes the system accessible to non-technical users without requiring programming skills. The modular architecture allows individual components to be updated independently without rebuilding the entire system. Ethical considerations including data privacy and prediction transparency were incorporated into the system design.

B. Limitations

Prediction accuracy is bounded by the quality and coverage of the training dataset; newly released models with unseen configurations may receive less accurate estimates. The current system does not incorporate dynamic market factors such as regional electricity tariffs, government incentive schedules, or supply chain disruptions. The ANN model's performance is constrained by dataset size; access to a larger corpus would likely improve its stability and close the performance gap with Random Forest. Predictions represent manufacturer suggested retail prices and may not reflect negotiated transaction prices or dealer markups.

X. FUTURE ENHANCEMENTS

Several directions would extend and strengthen this work. Live market API integration would allow continuous model retraining by connecting the system to real-time EV market data feeds, keeping predictions aligned with current pricing conditions. Deep learning architectures — such as Transformer-based models or gradient boosting variants including XGBoost and LightGBM — could be evaluated as higher-capacity alternatives to the current ensemble approach. Real-time resale value prediction incorporating depreciation curves and battery degradation models would extend the system's utility to total cost of ownership estimation. Integration of geospatial charging infrastructure data would enable the system to recommend optimal charging strategies alongside price estimates. Packaging the prediction interface as a cross-platform mobile application would broaden accessibility, particularly for consumers at dealerships. Expanding the feature set to include warranty duration, software update policy, and aftermarket maintenance cost data would improve prediction completeness.

XI. CONCLUSION

This paper presented a machine learning-based system for analyzing and predicting electric vehicle prices, motivated by the persistent challenge of cost transparency in a rapidly evolving market. Five supervised regression algorithms were evaluated on a structured dataset of EV specifications using standard metrics of MAE, RMSE, and R2.

Random Forest emerged as the most practically effective model, achieving an R2 of 0.9813 and an accuracy of 92.06% at a training time nearly twenty times faster than the ANN. The ANN achieved a marginally higher accuracy of 93.60% but at substantially greater computational cost and variance. Battery capacity, driving range, fast-charge speed, and market segment were identified as the most influential price determinants.

The deployed Django-based web system provides real-time, specification-driven price predictions accessible to consumers, manufacturers, and researchers. The results confirm that machine learning offers a credible, data-driven alternative to manual price estimation, with practical utility across market strategy, policy design, and consumer decision support. As EV fleets grow and datasets expand, the accuracy and scope of such systems will only improve — making data-driven cost transparency an increasingly achievable goal for the transition to sustainable mobility.

REFERENCES

[1] European Environment Agency, 'Electric Vehicles and the Energy Sector - Impacts on Europe's Future Emissions,' EEA Technical Report, 2016.
[2] S. Thangavel, M. Deepak, T. Girijaprasanna, S. Raju, C. Dhanamjayulu, and S. M. Muyeen, 'A Comprehensive Review on Electric Vehicle: Battery Management System, Charging Station, Traction Motors,' IEEE Access, 2023.
[3] D. Qiu, Y. Wang, W. Hua, and G. Strbac, 'Reinforcement learning for electric vehicle applications in power systems: A critical review,' Renewable and Sustainable Energy Reviews, vol. 173, p. 113052, 2023.
[4] C. Douillard and S. Audette, 'Comparaison des couts totaux de possession de vehicules electriques et conventionnels au Quebec,' Research Report, 2020.
[5] T. U. Solanke et al., 'A review of strategic charging-discharging control of grid-connected electric vehicles,' Journal of Energy Storage, vol. 28, p. 101193, 2020.
[6] A. Desreveaux, E. Hittinger, A. Bouscayrol, E. Castex, and G. M. Sirbu, 'Techno-economic comparison of the total cost of ownership of electric and diesel vehicles,' IEEE Access, vol. 8, pp. 195752-195762, 2020.
[7] H. L. Breetz and D. Salon, 'Do electric vehicles need subsidies? Ownership costs for conventional, hybrid, and electric vehicles in 14 US cities,' Energy Policy, vol. 120, pp. 238-249, 2018.
[8] N. B. G. Brinkel et al., 'Should we reinforce the grid? Cost and emission optimization of electric vehicle charging under different transformer limits,' Applied Energy, vol. 276, p. 115285, 2020.
[9] A. Van Velzen, J. A. Annema, G. van de Kaa, and B. van Wee, 'Proposing a more comprehensive future total cost of ownership estimation framework for electric vehicles,' Energy Policy, vol. 129, pp. 1034-1046, 2019.
[10] N. Lutsey and M. Nicholas, 'Update on electric vehicle costs in the United States through 2030,' International Council on Clean Transportation, 2019.
[11] Microsoft Azure Machine Learning Documentation, 'How to select algorithms,' https://docs.microsoft.com/azure/machinelearning/how-to-select-algorithms, accessed June 2023.
[12] V. Jatana, 'Machine Learning Algorithms,' SRM Institute of Science and Technology, Technical Report, June 2019.