



Machine Learning and Deep Learning Approaches for Global Electric Vehicle Sales Forecasting Under Low-Resolution Data Constraints

Abhishek Kumar^{*1}, Devesh Kumar Upadhyay¹ and Partha Sarathi Bishnu¹

¹ CSE, Birla Institute of Technology, Mesra, Ranchi, INDIA

^{*}*abhishek.atom@gmail.com, psbishnu@bitmesra.ac.in,*

deveshupadhyay3@gmail.com

ABSTRACT

Electric Vehicle (EV) sales are increasing globally. This has created need for accurate forecasting models for formulating policies and implementing industry decisions. This paper presents a comparative study of machine learning (XGBoost) and deep learning models (LSTM and Transformer) for EV sales forecasting on a multi-country, low-frequency dataset. Unlike prior works that use high-frequency or single-country data, this study focuses on yearly EV sales across multiple countries. Experimental results show that XGBoost together with careful feature engineering outperforms deep learning models, achieving an R^2 score of 0.99. LSTM and Transformer models achieve an R^2 score of 0.95 and 0.85 respectively in this study. The findings highlight the limitations of deep learning in small, low-resolution dataset.

Keywords: LSTM, Transformer, Gradient Boosting, Forecasting.

1. INTRODUCTION

The rapid growth of Electric Vehicles has not only reduced the carbon generated from fossil fuel driven vehicles but also created the need for developing better infrastructure for EV ecosystem. Thus, forecasting EV sales has become pertinent for timely development of EV infrastructure like electric charging stations, mining critical minerals for developing batteries used in electric vehicles, repair and servicing stations and others. Recent developments in technology have resulted in the increase in EV sales in the world altogether in the recent decade [1].

Forecasting EV sales has become important for policy makers, manufacturers and investors in recent years. This shift towards sustainable transport has led to the reduction of pollution and the emission of greenhouse gases. International commitment towards green transportation is reflected in the Paris Agreement to cut carbon emission in the near future [2]. Governments driven by these commitments are focusing on providing incentives to the public to switch to EV. People's commitment towards cleaner greener and sustainable transport has also led to the increase in demand of EV's world over [3].

Forecasting of EV sales helps the policy makers in the development of monetary incentives for adopting EV as well as the development of a robust ecosystem to support EV in a timely manner. This forecasting also helps manufactures to fix supply chain of manufacturing EV's. Investors can benefit from these forecasting by making informed decisions about their investments.

Prior work in EV forecasting depended on high-frequency or single-country data, this study focuses on yearly EV sales across multiple countries. Prior work used ARIMA and econometric models, while recent studies explore machine learning and deep learning methods. Most existing work focuses on high-frequency or single-region datasets. Global EV datasets are often low-frequency (yearly) and heterogeneous across countries. This raises an important question: which modeling approach performs best under such constraints?

This paper aims to:

- Compare XGBoost, LSTM, and Transformer models
- Evaluate performance on a multi-country EV dataset
- Analyse the impact of data limitations on model performance

A comprehensive Related work is discussed later followed by the Methodology, Results, Discussion and Conclusion.

2. RELATED WORK

Prior work on EV sales forecasting used linear regression and expert-based models. These suffer from low adaptability and also could not handle nonlinear market dynamics. Current works that have applied ML techniques have improved prediction accuracy but still they suffer from real-time adaptability and resistance to noisy market disturbances.

To address these gaps, Teng et. al, [4] introduced adaptive optimized grey models together with optimization algorithms (such as PSO). This enables dynamic tuning of parameter and improves forecasting accuracy.

Cui et. al., [5] used Linear Regression, Random Forest, and Gradient Boosting on EV sales dataset. Ensemble models such as Random Forest and Gradient Boosting outperform Linear Regression in EV sales forecasting.

Veysi et. al, [6] proposed a hybrid framework. They used SARIMAX, Prophet, and XGBoost models for UK EV sales prediction. They used machine learning and understandable time-series forecasting. Mishra et. al, [7] proposed a hybrid framework LSTM-GBRT for improving EV sales forecasting. It combines temporal learning with ensemble regression techniques. The study compared the hybrid model with Linear Regression, SVR, ARIMA, LSTM, using RMSE and R² metrics.

Srinivasan [8] used ARIMA and LSTM models on historical EV sales data. The study demonstrates that LSTM outperforms ARIMA in forecasting EV sales.

Existing EV sales forecasting studies mainly focus on individual statistical or deep learning models such as ARIMA, LSTM, SARIMAX, Prophet, and hybrid LSTM-GBRT frameworks. These works perform generalized multi-model benchmarking on different datasets. This paper addresses these gaps by integrating comprehensive preprocessing, feature engineering, and analyzes multiple machine learning and DL forecasting models to achieve more accurate and scalable EV sales prediction.

3. METHODOLOGY

The dataset used in this study is available on Kaggle. The dataset records yearly EV sales in many different countries. The main objective of this paper is to forecast worldwide EV sales. We preprocess the dataset by filtering some important features. We then engineer some new features based on statistical data. This new dataset is then used for forecasting EV sales using machine learning and deep learning. The original dataset contains the following features as depicted in Table 1.

Table 1. Features in the original dataset

Column Name	Description
Region	The country or region where the sales was tracked
Category	Classification of the data
Parameter	“EV sales” contains the actual sales recorded
Mode	Type of EV like Busses Cars etc.
Powertrain	Technology used (BEV, PHEV etc.)
Unit	Measurement unit (“Vehicles” for actual sales, “percent” for market share)

Year	Year of sale
Value	Vehicles sold or market share as percentage

3.1 Data Cleaning

The dataset contains the “value” column which contains two values, sale of vehicles or the percentage of market share. The sale of vehicles is filtered and the percentage value is discarded.

This directly demonstrates:

- irrelevant parameters removed
- BEV/PHEV aggregated
- transformed into forecasting-ready format

Converted data into (country, year, sales) format for implementing forecasting.

3.2 Feature Engineering

In order to improve forecasting of EV sales, several temporal and statistical features are engineered from the cleaned EV sales dataset. These features capture historical trends, country-specific behaviour, and sales variations over time. The final dataset after feature engineering contains the following attributes:

Table 2: Final feature set used in the model

Feature	Description
Country	Name of the country or region associated with the EV sales record.
Year	Year corresponding to the recorded EV sales value.
Sales	Total EV sales for the given country and year after aggregating BEV and PHEV sales.
country_id	Numerical encoding assigned to each country to enable machine learning models to process categorical country information.
lag_1	EV sales from the previous year.
lag_2	EV sales from two years prior.
lag_3	EV sales from three years prior.
lag_4	EV sales from four years prior.
lag_5	EV sales from five years prior.
rolling_mean_3	Three-year rolling average of EV sales used to capture smoothed long-term trends.
rolling_std_3	Three-year rolling standard deviation representing short-term volatility and variability in sales patterns.
Growth	Logarithmic Year wise growth rate of EV sales

The lag-based features enabled the models to learn temporal dependencies from historical EV adoption patterns. Trend stability and market volatility across countries is captured using Rolling mean and rolling standard deviation [9]. The growth feature gives the logarithmic growth rate which captures the acceleration and slowdown of EV adoption across countries. The raw yearly sales data is augmented with these engineered features which helps in forecasting. The final feature set is then fed to machine learning and deep learning frameworks such as XGBoost, LSTM, and Transformer architectures.

Training uses all data except the final two years. The final two years data is used for testing. XGBoost, LSTM, and Transformer models were trained on the engineered augmented dataset for forecasting. The XGBoost model used gradient boosting regression trees used for optimization. LSTM and Transformer used sequential input windows for training. The LSTM and Transformer detect temporal dependencies in the data.

Evaluation of the models was done using:

- Mean Absolute Error (MAE)

- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)

These metrics were selected to measure prediction accuracy, error magnitude, and the overall explanatory capability of the forecasting models. Python was used as the coding language.

3.3 XGBoost

XGBoost (Extreme Gradient Boosting) [10] is an ensemble machine learning algorithm that utilizes gradient boosting over decision trees to achieve high performance. XGBoost iteratively constructs decision trees. Each decision tree minimizes the errors of the previous decision trees. Structured tabular datasets can benefit from ability of XGBoost to model nonlinear relationships, handle heterogeneous data distributions, and reduce overfitting. XGBoost leads the Transformer and LSTM models with highest R^2 score and lowest prediction error metrics. Thus, we can say that XGBoost is apt for heterogeneous EV sales data that is limited and highly variable.

3.4 LSTM

In this study, the LSTM model uses sequential historical sales data represented through lag-based time windows to forecast EV sales. The model learns temporal patterns and long-term growth trends from the augmented dataset. Due to its ability to process sequential information, LSTM is well-suited for time-series forecasting tasks. LSTM model produced smoother predictions and struggled to capture abrupt fluctuations in yearly EV sales data in our study. The limited size and low temporal dataset hampered model’s performance to learn the complexity in the heterogeneous dataset.

3.5 Transformer

Transformers use self-attention mechanism to identify important temporal relationships and changes in EV adoption patterns across different years. The Transformer model captures the sudden spikes or fluctuations in EV sales more effectively as compared to LSTM. Transformer architectures generally require large-scale datasets to exploit attention-based learning, The model produced smoother predictions and was less effective at capturing extreme volatility in certain regions. This behaviour can be attributed to the low frequency heterogeneous EV sales dataset. The Transformer provided valuable insights into the applicability of attention-based deep learning methods for global EV sales forecasting.

4. RESULTS

4.1 Quantitative Results

Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) were used to evaluate the performance of XGBoost, LSTM, and Transformer models. Table 3 presents the results for all 3 models.

Table 3: Model Performance

Model	MAE	RMSE	R^2
XGBoost	0.24	0.38	0.99
LSTM	0.43	0.60	0.95
Transformer	0.80	1.06	0.85

The results indicate that XGBoost outperforms LSTM and Transformer in the forecasting task. It not only has the lowest MAE and RMSE but also the highest R^2 score. Its R^2 value of 0.99 shows that it could explain the variance in EV sales with 99% accuracy. XGBoost can learn nonlinear relationships from the engineered temporal features which explains its better performance. The LSTM beats the Transformer model in terms of MAE, RMSE and R^2 values. Figure 1and Figure 2 shows the comparison of all three models.

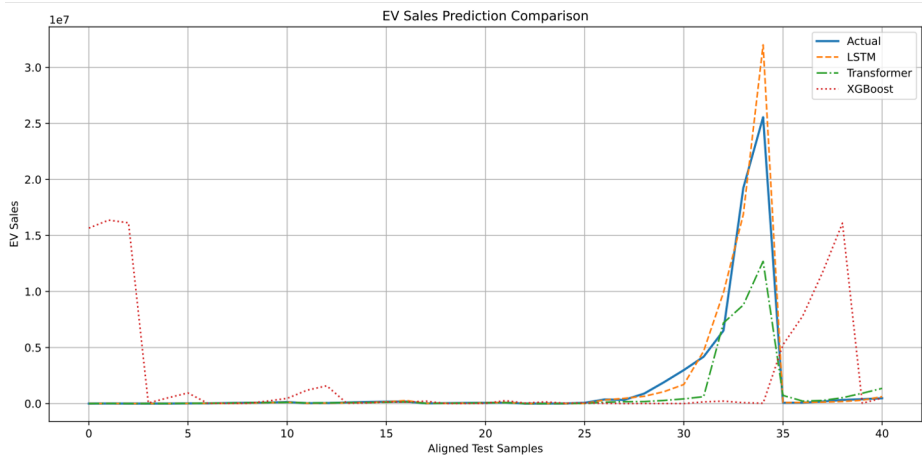


Figure 1. EV Sales Prediction Comparison

5. DISCUSSION

5.1 Comparative Analysis of Models

The superior performance of XGBoost can be attributed to its effectiveness in using structured tabular features such as lag values, rolling statistics, and growth indicators. The dataset included yearly EV sales data with limited historical records. Therefore, the machine learning methods worked better than deep learning models that need more data.

The LSTM model captured the general trend of EV adoption. It failed to accurately predict sudden fluctuations and rapid market changes. This may be due to the limited sequential information available in the yearly dataset.

The self-attention mechanism of Transformer was able to predict sudden growth better than LSTM. The Transformer performed better than the LSTM in predicting the sudden growth patterns in EV sales. However, it could not accurately show the large variations in the EV market. Figure 2 shows the log-scale plots which reiterates these findings.

5.2 Impact of Dataset Characteristics

The highly heterogeneous nature of the global EV sales dataset that included data from countries with vastly different market sizes and adoption rates, resulted in substantial variance across observations. Large EV markets such as China and the United States produced significantly higher sales values than smaller emerging markets. This led to scale imbalance and strong non stationarity in the forecasting problem.

Logarithmic scaling and temporal feature engineering were employed to address the above challenges. The use of log-scale visualization (Figure 2) provides the prediction behaviour across both low-sales and high-sales regions. XGBoost could capture nonlinear variations and abrupt market changes in the experiments. The DL models Transformer and LSTM produced smoother predictions in the same experiments.

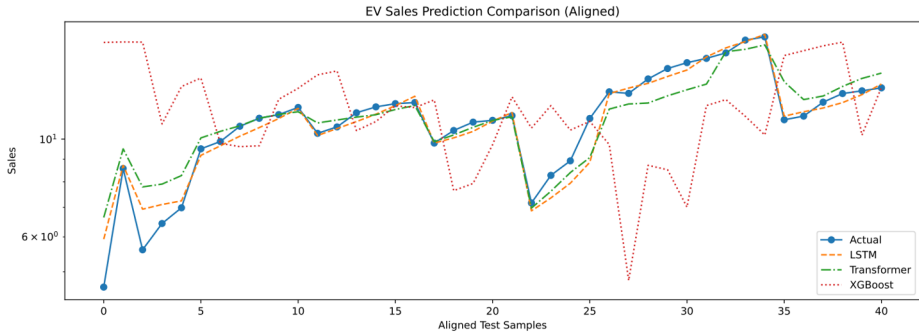


Figure 2. EV Sales Prediction in log scale.

6. CONCLUSION

The dataset used in the study was heterogeneous and of limited size. This handicaps Transformer and LSTM in forecasting task. DL models like Transformer and LSTM rely on high frequency datasets to exploit temporal relationships. XGBoost with R^2 value of 0.99 and low MAE and RMSE was able to beat these complex models. This also proves the point that XGBoost together with careful feature engineering can outperform complex DL models in low frequency datasets. This also proves that for real-world energy and transportation forecasting problems with limited historical observations, traditional ML models may be better than more complex DL models. Future work may be done on other similar limited frequency datasets.

Acknowledgment: None

7. REFERENCES

- [1] X. Long, C. Liu, B. Miao, J. Xiao, Y. Cao, Z. Xu, C. Li, L. Wang, and Q. Yang, "Life Cycle Assessment of Typical New Energy Vehicles in China," *J. Phys. Conf. Ser.*, vol. 3000, no. 1, pp. 1–9, 2025, doi: 10.1088/1742-6596/3000/1/012016.
- [2] D. Nasreddin, H. E. Hafdaoui, F. Jelti, A. Boumelha, and A. Khallaayoun, "Inhibitors of Battery Electric Vehicle Adoption in Morocco," *World Electr. Veh. J.*, vol. 15, no. 1, pp. 6, 2023, doi: 10.3390/wevj15010006.
- [3] E. Javanmardi, M. Hoque, A. Tauheed, and M. Umar, "Evaluating the Factors Affecting Electric Vehicles Adoption Considering the Sustainable Development Level," *World Electr. Veh. J.*, vol. 14, no. 5, pp. 1–19, 2023, doi: 10.3390/wevj14050120.
- [4] Teng, Fei. "Beyond traditional forecasting: Machine learning and adaptive algorithms in EV sales predictions." *Applied and Computational Engineering* 82 (2024): 12–16.
- [5] Cui, Christopher Genzhuo. "Forecasting Future EV Sales: A Comparative Study of Model Performance." In *Proc. 2nd Int. Conf. on Data Analysis and Machine Learning (DAML 2024)*, pp. 75–79. 2024.
- [6] Veysi, Shima, Mohammad Moshfeghi, Amir Sadrfaridpour, and Peiman Emamy. "Electric Vehicle Sales Forecast for the UK: Integrating Machine Learning, Time Series Models, and Global Trends." *Algorithms* 18, no. 7 (2025): 430.

- [7] Mishra, Rishav Dev, Santanu Kumar Dash, Saichol Chudjuarjeen, Satyajit Mohanty, Shivam Prakash Gautam, Rupali Mohanty, and Md Tasnin Tanvir. "Hybrid LSTM-GBRT based machine learning technique implementation for electric vehicle sales prediction analysis." *International Journal of Sustainable Engineering* 17, no. 1 (2024): 843-858.
- [8] Srinivasan, Bhavana. "Predicting Future Electric Vehicle (EV) Sales: A Time Series Forecasting Approach Using Historical EV Sales Data." *International Journal for Applied Information Management* 5, no. 3 (2025): 177-189.
- [9] Upadhyay, D.K., Mohapatra, S. & Singh, N.K. (2023). An early assessment of Persistent Depression Disorder using machine learning algorithm. *Multimedia Tools Applications*. <https://doi.org/10.1007/s11042-023-17369-4>
- [10] Upadhyay, D. K., Mohapatra, S., & Singh, N. K. (2024). An early assessment of subjective well-being in an undergraduate by using heterogeneous stacked ensemble. *AIP Conference Proceedings*, 3164(1), 040001. <https://doi.org/10.1063/5.0214163>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

