IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AN ENSEMBLE APPROACH FOR THE PREDICTION OF SOC CONSUMPTION IN ELECTRIC VEHICLES

[1]S. Senthurya, [2]Dr. R. Senthamil selvi
[1]PG student, [2]Associate Professor
Department of Computer Science & Engineering,
Saranathan College of Engineering, Trichy, India.

Abstract: The rising acceptance of electric vehicles in today's world has brought a spotlight on how vital it is to precisely compute the state of charge (SOC) for efficient usage of battery capacity and vehicle recharging schedule for lengthy travel. This paper proposed a new stacking ensemble approach of two machine learning algorithms: XGBoost regressor and random forest. This works on the benefits of both algorithms to increase the model's prediction accuracy. It takes several attributes such as distance, speed, driving condition, and altimetry of the route into account to make accurate SOC estimation. The model consists of two stages, a base model and a meta-model. The base model gives the prediction of the XGBoost regressor and Random Forest individually and the meta-model gives the overall prediction. Finally, the result obtained from the experiment shows that the SOC prediction of the ensemble model is better than the individual models.

IndexTerms: **State of Charge (SOC), XGBoost regressor, Random Forest, Stacking, Electric Vehicles**

I.INTRODUCTION:

An electric vehicle battery or traction battery is a secondary rechargeable battery that powers an electric motor in both vehicles that run on electricity, such as hybrid and battery electric vehicles. The most affordable and feasible technologies available today are Lithium-ion batteries with a high specific capacity versus weight ratio. Integration of such Lithium-Ion Battery technology plays a vital role in modern electricity infrastructure like smart grids for load leveling, peak shaving techniques, and frequency regulation while reducing network interruptions [1]. Electrical fires resulting from malfunctions mostly occur with poor-quality batteries, therefore developing highly collaborative monitoring systems to identify any internal faults or degradation processes on time should be a priority [2]. Though other rechargeable batteries have been developed by manufacturers over the years to overcome lithium-ion technology limitations such as its limited lifespan – they remain less popular.

The most commonly used lates t battery types are Valve Regulated Lead Acid (VRLA), NiCd, NiMH, Zinc Air, and Sodium Nickel Chloride ("Zebra") batteries. Other power storage device technologies that offer higher power density than all these battery types are Supercapacitors but their ability to store energy is low compared to lithium-ion batteries. Electric vehicle functionality improves significantly by bringing together both these technological innovations [3]. To ensure an efficient run, keeping an eye on the State of Charge (SoC) proves critical as it denotes how much energy remains in the battery till it needs a recharge. Electric vehicle's reliable SOC consumption is important in augmenting vehicle performance, and effective energy management systems application while enhancing range estimates accuracy. Several approaches exist for determining SOC usage in electric cars, with commonly used ones being support vector regression or neural networks-based machine learning algorithms.

A hybrid model fusing deep learning models with Gaussian process regression improves SOC prediction results by combining different modeling techniques seamlessly [5]. Furthermore, Hybrid Ensemble Data-Driven method integrates two random learning algorithms that uncover associations between health indicators and practical SOH allowing greater accuracy when predicting Li-ion batterie's state of health, along with remaining useful life [6]. In contrast to traditional methods relying mostly on simplistic models or limited data input availability; deep learning approaches leverage contextual driving data temporal dependencies making them more efficient in improving SOC prediction accuracy alike an LSTM-based model suitable for producing precise short-term battery state projections useful for effective energy management implementation resulting from better range estimation, enhanced EV performance optimization [7,8].

To optimize energy usage, improve navigation, and provide a better user experience, various machine-learning techniques have been employed in models predicting the state of charge (SOC) depletion of electric vehicles [15]. Among these techniques is the ensemble method which combines multiple models to yield improved prediction accuracy and tolerance. Leveraging the diverse intelligence of individual models, ensemble methods achieve more robust predictions. XGBoost and Random Forest are two popular algorithms employed in SOC prediction through ensembles. Their demonstrated success across varied domains and ability to handle complex relationships in EV SOC consumption makes them well-suited for this application.

II.RELATED WORK:

A machine learning model designed by J.P. Ortiz et al. utilizes empirical data from EVs to project State of Charge (SOC) consumption with precision leveraging continual reinforcement learning techniques alongside meta-experience replay techniques and employing an artificial neural network as its foundation. As opposed to traditional models needing re-working from scratch every time requirements change, this one offers the added benefit of ongoing discovery and adjustment continuously while incorporating new insights from real-time feedback mechanisms using an existing EV fleet's available datasets for training purposes. Nonetheless, since these models may tend towards overfitting bias with heavy reliance upon prevailing patterns of historical usage there exists a potential drawback where the generalized application could result in less dependable outcomes producing an accuracy of 91 %.

Bin Gou et al. used the hybrid data-driven model to predict the SOC and RUL values. The model is a combination of ELM and RVFL to map the health indicator with the practical soc. It has two types of working processes such as online and offline prediction. The online prediction uses the NARX structure which is used for the SOH prediction. And for the RUL prediction based on the NAR ensemble learning and bootstrap mechanism. This model provides better results by using various ensembles and learning techniques.

Mona Faraji Niri et al. focus on lithium-ion batteries to predict the state of available power (SOAP). To examine the dynamics of the battery, the ECM is first implemented and then incorporated with the model. This model is designed with Wavelet analysis and Markov model a long-term load prediction model is created. The SOAP analysis is then put together with sentiment analysis to provide better results. Z.Lyu et al. introduced a fusion method of the metabolic grey model and multiple-output Gaussian process regression. In addition to this battery aging investigations are carried out to check the working of the fusion method. And thus provides the SOH prediction.

Jing Li et al. introduced a neural network model with the combination of Bidirectional long short-term memory (BiLSTM) and recurrent neural network (RNN) to get accurate SOC prediction. The advantage of using a neural network is that the model can move forward and backward for better prediction results.

III.METHODOLOGY:

3.1) Data Collection and Preprocessing

Collecting accurate data is an integral part of research and analysis projects as it provides raw material to analyze trends, patterns, and correlations between various aspects discussed. The precision & quality of the collected data holds paramount importance when drawing persuasive conclusions from the analysis performed. This study utilized datasets by collecting two different types- The first one consisted of about 1000 records sourced from KAGGLE containing essential details like source and destination of latitudes and longitudes; speed, distance traveled through various routes; travel time alongside elevation gained during each route traversed etc., rendering it useful for state-of-charge (SOC) prediction related projections. On the other hand, the second dataset for this study involved around 3346 records procured via CALCE providing crucial distance, speed, and travel time alongside diverse driving styles and tire type applications affiliated with SOC consumption information rendered via ECR derivation techniques.

Raw data can be inconsistent and inaccurate; hence it requires proper preparation for appropriate analysis through what we refer to as pre-processing techniques. One effective way to preprocess your dataset is by uploading it onto a pandas DataFrame and making use of pre-existing functions like dropna() or fillna() for coping with any missing values in case you are working on Google Colab's platform. The outliers can be spotted using statistical approaches like the z score or interquartile range (IQR) and can be dealt with either by deleting them entirely or by replacing them with the appropriate values.

z-score standardization:

z = (x - mean) / standard deviation, where 'x' represents the feature value.

And the next point to consider is, converting the categorical values into numerical values via encoding techniques namely one-hot encoding or label encoding which depends upon the nature of variables and the requirement of the model. Feature Selection requires that we identify the most important features for SoC prediction; these can be achieved using the correlation analysis (corr()) or feature importance analysis (feature_importances_) in the sklearn library, with threshold values specified by correlation coefficient or feature importance score being used to select essential features only. Irrelevant and redundant features should be removed from datasets to improve model efficiency and minimize over-fitting. Also, Feature Scaling techniques such as StandardScaler() or MinMaxScaler() in the sklearn library can be used to scale the data attributes appropriately. to identify if there is any strong relationship between the features, the pairwise correlation is analyzed. If a high correlation between the features is identified it may cause redundancy and lead to multicollinearity issues. It is best to use correlation matrics or scatter points to observe the relationship. Principal Component Analysis (PCA) is used to handle the high dimensionality data which transforms the features into uncorrelated variables namely principal components. The amount of variance in the data helps those principal components to be ordered. without losing the information, the dimensionality can be reduced by selecting the subset of principal components. Both the preprocessed datasets are split into training and testing datasets of the ratio 80% and 20% using sklearn's train_test_split() function.

3.2) SOC Prediction training using Machine learning

In this study, two machine learning algorithms namely XGBoost regressor and random forest were ensembled together to predict the SOC consumption of electric vehicles.

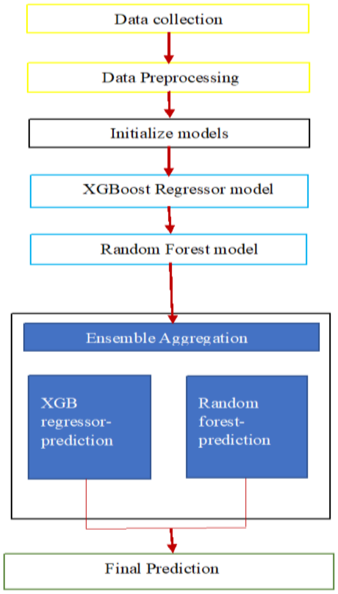


Figure 1: Workflow of the SOC Prediction Model

The project's overall structure is to feed the preprocessed dataset into the models to learn the patterns and relationships that will predict the accurate prediction of soc consumption.

3.3) Implementation of XGBoost regressor

XGBoost regressor is a powerful gradient-boosting algorithm that presents greater results in various prediction tasks. It works on the principle of a combination of various decision trees which creates a strong ensemble model.

Table 1: Hyperparameters of XGBoost regressor model

Hyperparameters	*Specifications*
Base score	0.5
Booster	Gbtree
Learning Rate	0.3
Maximum depth	6
Number of estimators	100
colsample_bylevel, colsample_bynode, gamma, min_child_weight, reg_alpha, reg_lambda, scale_pos_weight, tree_method, subsample	Default

The XGBoost regressor model is initialized with specific hyperparameters as mentioned in table 1 which are tuned carefully to get optimized soc prediction.

The XGBoost regressor model was trained on both the preprocessed dataset by building the decision trees sequentially. each tree is trained to capture the residuals from the previous tree. In both the datasets (1 & 2) soc consumption is chosen as a target variable and other need features as input. This makes it easier for the model to observe the relationship between the soc and other input features. This helps the model to focus on the area where the previous tree lacks and helps the model to perform well. To find the variance between the actual and expected soc values MSE is calculated which gives the average squared difference.

MSE = (1/n) * Σ(y_p - y_a)^2

where n is the number of samples, y_p represents the predicted SoC values, and y_a represents the actual SoC values.

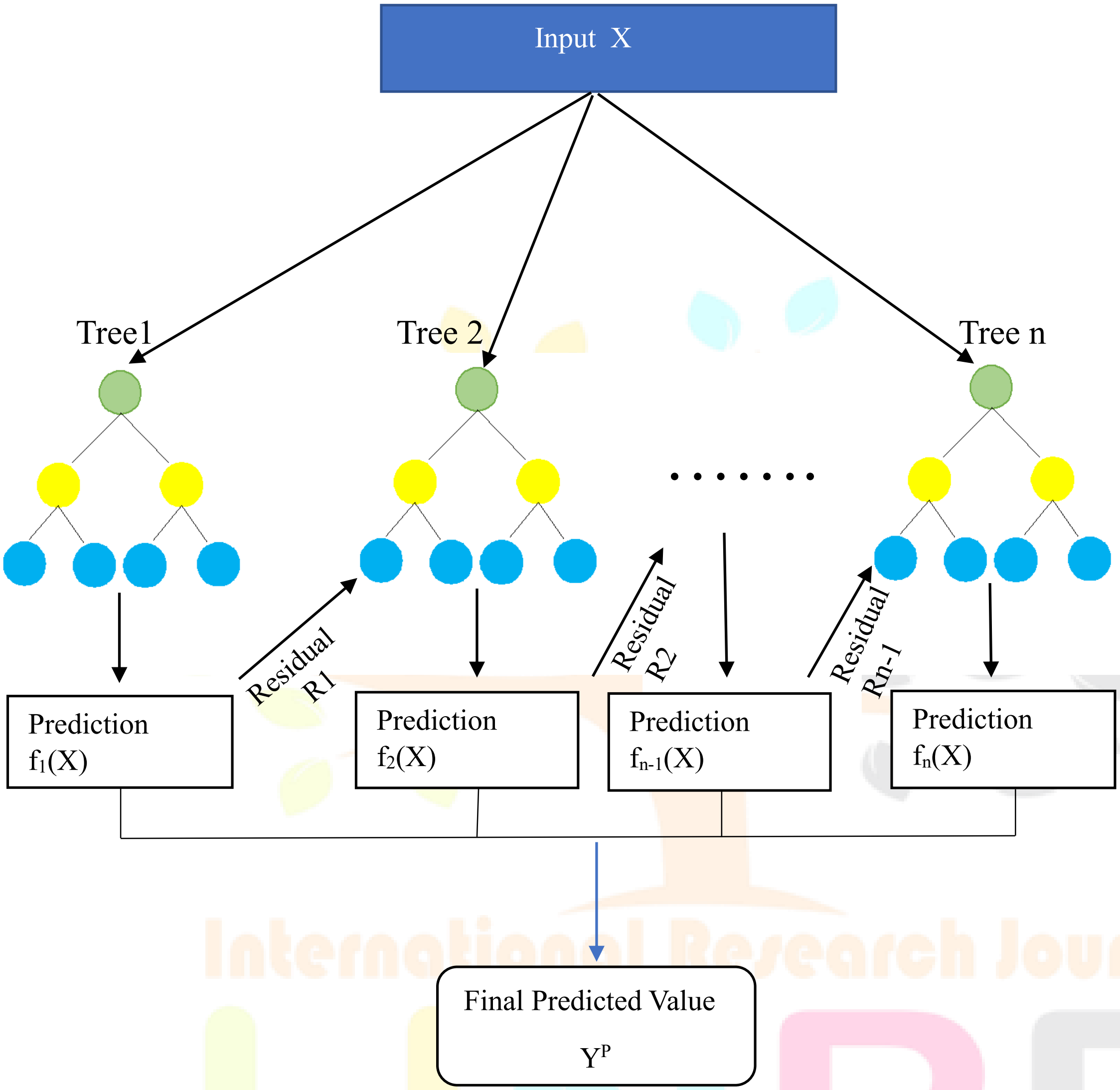


Figure 2: Soc prediction structure of XGBoost regressor

International Research Journal IJNRD Research Through Innovation

It reduces the objective function which provides the variance between actual and predicted soc values. Hence, it is estimated using second-order Taylor expansion. Then the gradients are calculated to update the prediction using gradient descent optimization. The final soc prediction is obtained by combining all the individual trees.

The objective function can be determined using the below formula

$$l(y_i, y_p)= (y_i - y_p)^2$$

$$obj = \Sigma[l(y_i ,y_p) + \Omega(Y^P)]+ \gamma T \qquad (1)$$

Where y_i and y_p are the actual and expected soc values of the individual values, $l(y_i,y_p)$ is the loss function, $\Omega(Y^P)$ is the regularization term, γ is the regularization parameter to reduce complexity and T is the number of trees.

The initial prediction y_o is the average of the target variable y, to start the iterations it works as the starting point,

$$y_o = mean(y) \qquad (2)$$

The model starts constructing the recursion tree iteratively to form an ensemble of decision tree and in each iteration, new residuals are calculated which represents the error in the previous iteration made by the model,

$$r = y_{i\,-}\, y_{n-1} \qquad (3)$$

where y_{n-1} is the prediction made in the current iteration. The new decision trees are started building according to the residuals. Then for each data point, the gradient(gi) and hessian(h_i) are calculated by,

$$g_i = \partial l(y_i,\ y_{n-1})/\partial y_{n-1} \tag{4}$$

$$h_i = \partial^2 l(y_i,\ y_{n-1}) \ /\ \partial (y_{n-1})^2 \tag{5}$$

The gradient and hessian help to find the decision tree's ideal structure by iteratively splitting the data based on different features.

The regularization term is calculated by the following formula,

$$\Omega(\bar{y}) = \gamma T + \tfrac{1}{2}\lambda \sum [g^2 \ /\ (h + \lambda)] \tag{6}$$

Where λ is the regularization parameter controlling the amount of shrinkage applied to each tree, g is the gradient of the loss function, indicating the direction to reduce the objective function, and h is the Hessian, describing the curvature of the objective function. The objection function is updated after adding a new tree,

$$\mathrm{Obj}_t = \sum [l(y_i,\ \bar{y}_{t-1} + f_t(x_i))] + \Omega(\bar{y}_t) \tag{7}$$

The final prediction of the soc values,

$$Y^P(x) = y_0 + \eta * \sum_n f_n(x) \tag{8}$$

Where $Y^P(x)$ represents the final prediction for the input sample x, y_0 is the initial prediction, and η is the learning rate. Finally, the XGBoost regressor model combines the strengths of gradient boosting, tree splitting, and learning rate to get accurate SOC prediction.

3.4) Implementation of Random Forest

Random forest is a collective learning algorithm that makes use of various decision trees to make accurate forecasts. The model work properly when it utilizes random subsampling of features and the hyperparameters are tuned accordingly.

Random forest prediction model is trained on both datasets 1 and 2, having the state of charge feature as the desired parameter. Considering the chosen characteristics, the n number of decision trees are built and that helps to capture the relationship and pattern of the data. Then by combining the decision trees through averaging ensemble technique to make the final soc prediction.

Table 2: Hyperparameters of the Random Forest model

Hyperparameters	*Specifications*
Number of estimators	100
Maximum depth	None
Criterion, min_samples_split, min_samples_leaf, max_features, max_leaf_nodes	Default

The final SOC prediction of an electric vehicle can be calculated in the random forest model using the averaging technique,

$$\text{Final Prediction} = (1/T) * \Sigma y_p \tag{9}$$

Where T is the number of trees and y_p is the soc prediction of the individual decision tree.

IJNRD
Research Through Innovation

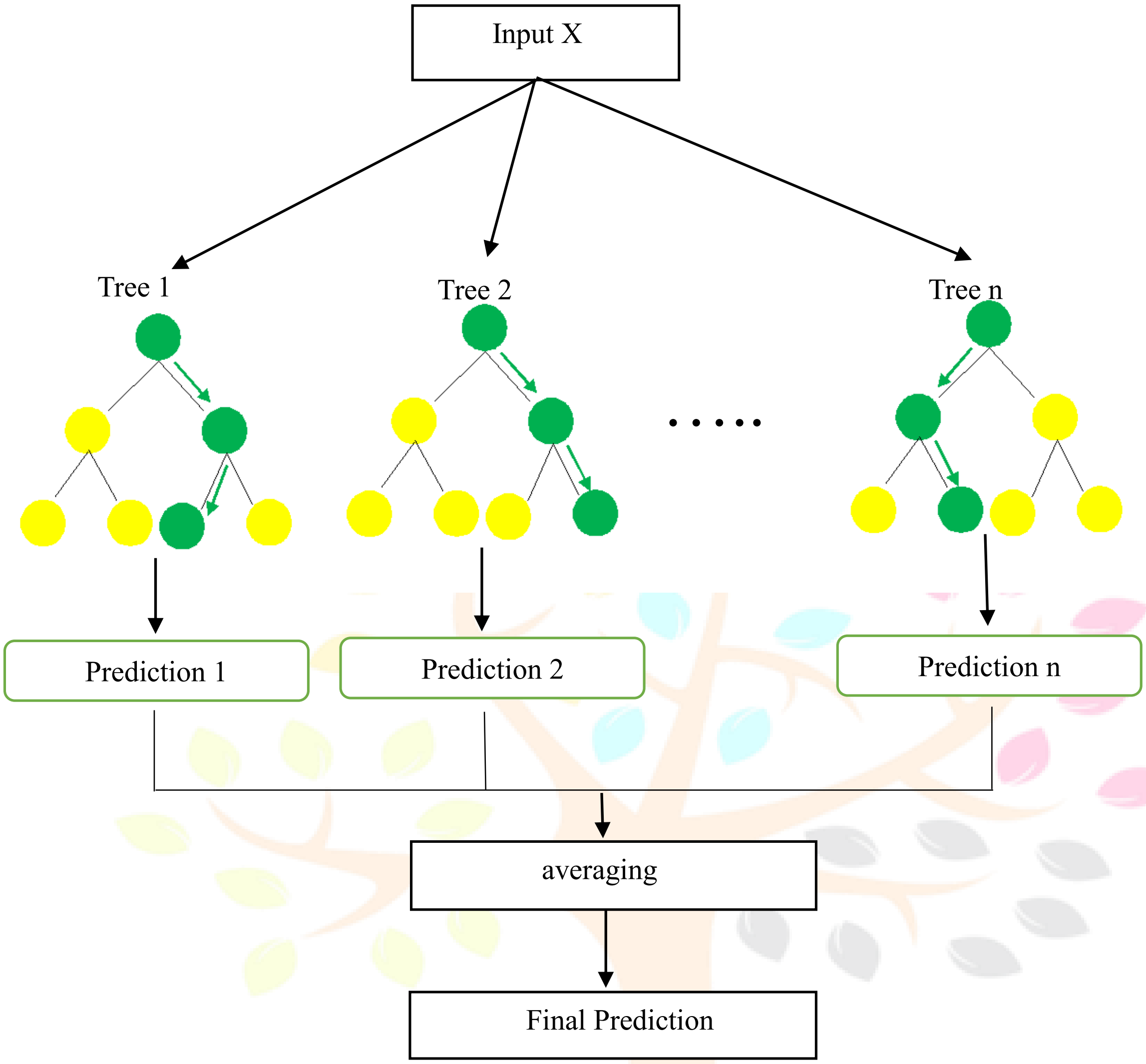


Figure 3: SOC prediction structure of Random Forest

3.5) Working of the Ensemble model: Stacking

Two different models are trained with two different datasets for the prediction of soc values, to further improve the accuracy of the model and the performance an ensemble technique called stacking is used. The input for this ensemble model is the SOC prediction calculated by both the XGBoost regressor and the Random Forest.

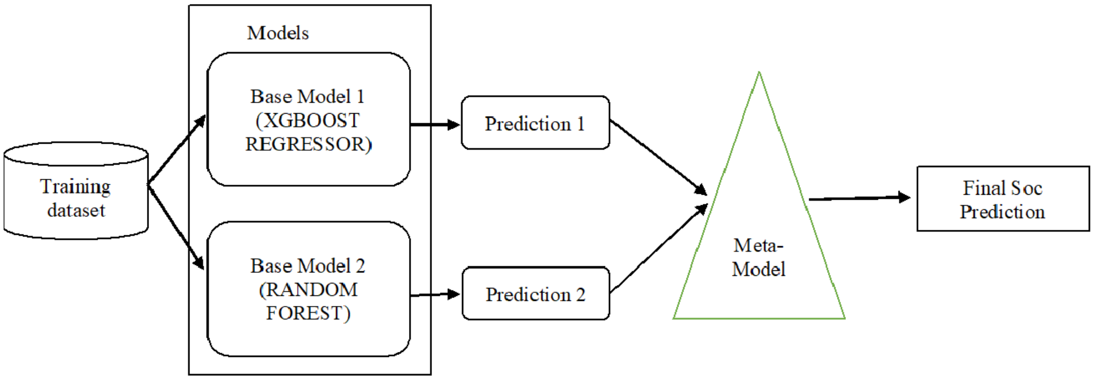


Figure 4: Working structure of the ensemble model

The Stacking ensemble technique creates a Meta-Model which takes the prediction of the models as the input. The meta-model used here is the linear regression model. To make the final SOC prediction by the Stacking ensemble technique,

$$Y\_final = \beta_0 + \beta_1 * y\_base1 + \beta_2 * y\_base2 \qquad (10)$$

Where y_base1 is the soc predicted of the first base model (XGBoost Regressor), y_base2 is the soc predicted value of the second base model (Random Forest), β_0, β_1, and β_2 are the coefficients of the linear regression model.

IV.RESULT AND DISCUSSION:

This Model is trained and tested using two different datasets 1 and 2 of 1000 and 3346 records respectively with different attributes but both have the target variable as SOC. The model is then evaluated using the R-Squared score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These values help to better understand the model and also the performance of the model. The R-Squared (R2) score can be calculated by,

$$R2 = 1 - (\Sigma(y\_p - y\_a)^2 / \Sigma(y\_a - y\_mean)^2) \qquad (11)$$

Where y_mean is the mean of the target values.

The performance of individual models such as XGBoost Regressor, Random Forest, and LightGBM are analyzed and compared. Then the efficiency of various ensemble models are analyzed and the model that provides better result for this soc prediction model is identified.

4.1) Performance of Individual Models

First XGBoost Regressor, Random Forest, and LightGBM algorithm are trained and tested individually to note the performance of each model. XGBoost regressor model trained by both datasets and gave good results in both datasets. It showed an accuracy of 92% and 82% respectively. Next random forest model was trained by both datasets and produced an accuracy of 91% and 82%. Finally, the LightGBM gives accuracy of 89% and 81% respectively.

Table 3: Individual model performance of Dataset 1

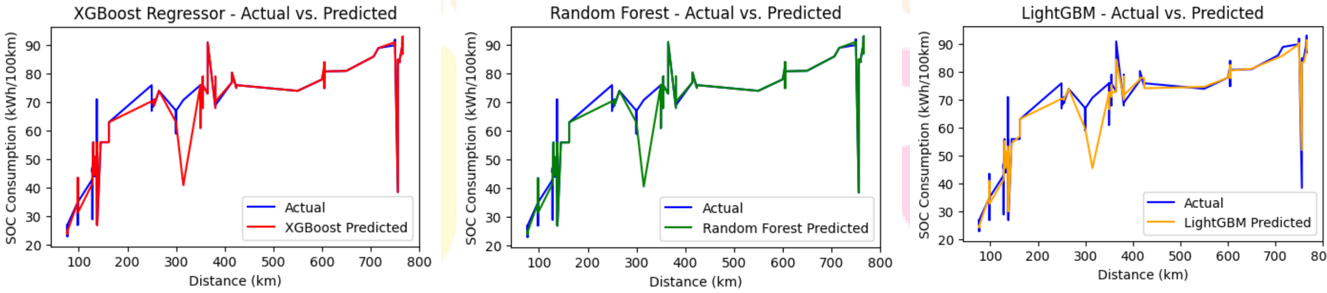
	R^2	MSE	RMSE
XGB Regressor	0.9220	24.6529	4.9652
Random Forest	0.9120	24.6587	4.9658
Light GBM	0.8905	29.5716	5.4380

Dataset 1 showed better results for all the models compared to Dataset 2. And by considering both the dataset's performance XGBoost regressor showed better accuracy.

Table 4: Individual model performance of Dataset 2

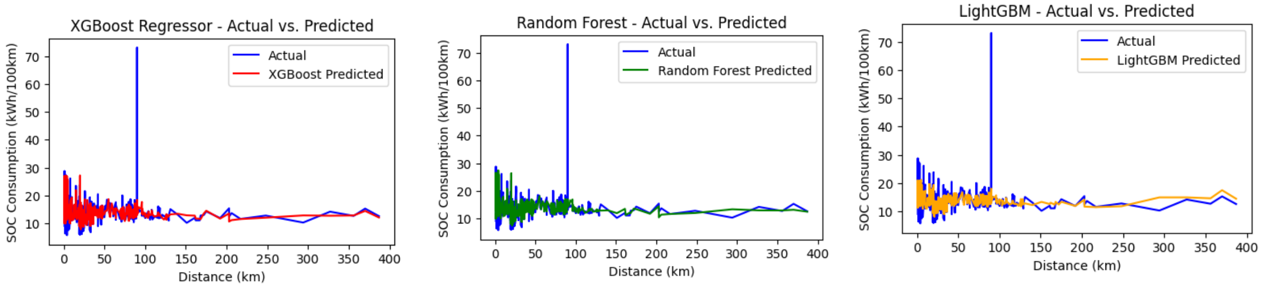
	R^2	MSE	RMSE
XGB Regressor	0.8293	8.0039	2.8291
Random Forest	0.8212	8.1409	2.8532
LightGBM	0.8138	11.3264	3.3655

Thus, Tables 3 and 4 show the performance of both the datasets to the respective models. The actual and predicted relationship between the soc consumption and distance is shown below.



a) b) c)

Figure 5: Actual and predicted soc consumption of dataset 1 in a)XGBoost regressor b)Random forest c)LightGBM



a) b) c)

Figure 6: Actual and predicted soc consumption of dataset 2 in a) XGBoost regressor b) Random Forest c) LightGBM

4.2) Performance of Ensemble Models

To further increase the precision of the soc estimation model several ensemble techniques such as averaging and stacking are considered. These two techniques are used to combine different individual models and provide various ensemble models. And based on the performance that gives greater accuracy is considered as the final model.

The ensemble model of XGBoost regression, Random Forest, and LightGBM through averaging produced an accuracy of 90% and 82% for both datasets 1 and 2 respectively. Then the ensemble model of XGBoost regression and Random Forest through averaging produced 91% and 82% for both datasets respectively.

The ensemble model of XGBoost regression, Random Forest, and LightGBM through stacking produced an accuracy of 93% and 83% for both datasets 1 and 2 respectively. Then the ensemble model of XGBoost regression and Random Forest through stacking produced 94% and 85% for both datasets respectively.

Table 5: Ensemble model performance of dataset 1

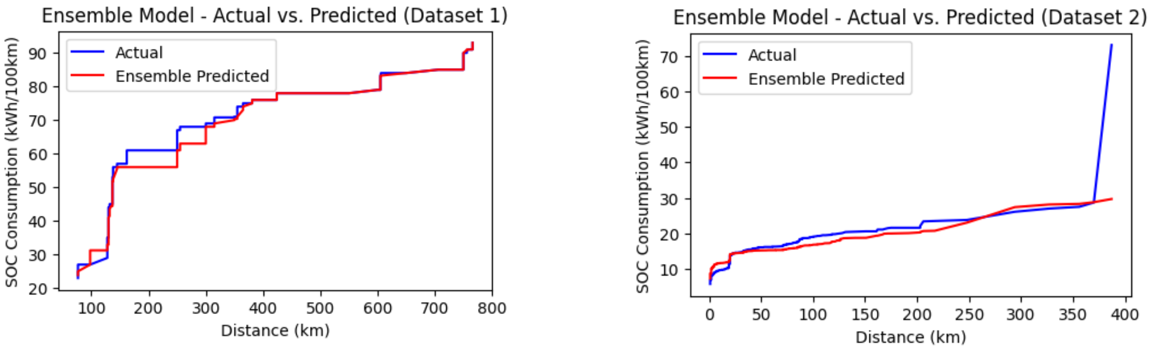
	R^2	MSE	RMSE
Ensemble method- Averaging(XGBoost regressor, Random forest, LightGBM)	0.9081	26.2944	5.123
Ensemble method- Averaging(XGBoost regressor, Random forest)	0.9170	24.6558	4.9655
Ensemble method- Stacking(XGBoost regressor, Random forest, LightGBM)	0.9355	27.6980	5.2629
Ensemble method- Stacking(XGBoost regressor, Random forest)	0.9420	24.6529	4.9652

Table 5: Ensemble model performance of dataset 1

	R^2	MSE	RMSE
Ensemble method- Averaging(XGBoost regressor, Random forest, LightGBM)	0.8214	9.3257	3.0159
Ensemble method- Averaging(XGBoost regressor, Random forest)	0.8252	8.0724	2.8411
Ensemble method- Stacking(XGBoost regressor, Random forest, LightGBM)	0.8315	7.6181	2.7601
Ensemble method- Voting(XGBoost regressor, Random forest)	0.8557	7.7346	2.7811

By considering all the performances of the individual models and ensemble models, the ensemble model of XGBoost regressor and Random Forest provides greater accuracy of all and suits well for the prediction of soc consumption. The actual and predicted soc values are shown in below to understand the relationship and prediction.

International Research Journal
IJNRD
Research Through Innovation



a) b)

Figure 7: Actual and prediction SOC value of the final ensemble model a)Dataset 1 b)Dataset 2

V.CONCLUSION AND FUTURE WORK:

In this study, it is presented that the performance of the ensemble model of XGBoost regressor and Random Forest is the best of all the other explored models for the prediction of SOC consumption in electric vehicles. And Dataset 1 and Dataset 2 performed well on all those machine-learning models but Dataset 1 performed well on the models compared to Dataset 2. In this, both the stacking and averaging techniques are employed and stacking is considered as the final one. In the Future, continual learning is planned to be added to this study. Then the continual learning learns from the model itself and makes the real-life correction to the model and then retrains the model according to it. Thus, this provides better results than the normal model. In addition to this can see different models or fine-tune the parameters and also add charging stations available in the specific route.

REFERENCES:

[1] J. P.Ortiz, G. P.Ayabaca, A. R.Cardenas, D.Cabrera, and J. D. Valladolid, "Continual Reinforcement Learning Using Real World Data for Intelligent Prediction of SOC Consumption in Electric Vehicles", IEEE Latin America Transactions, Volume 20,2022.

[2] Carkhuff, B.G., Demirev, P.A. and Srinivasan. Impedance-Based Battery Management System for Safety Monitoring of Lithium-Ion Batteries. IEEE Transactions on Industrial Electronics, Volume 65, 2018

[3] Carter, R., Cruden, A. and Hall, P.J. Optimizing for Efficiency or Battery Life in a Battery/Supercapacitor Electric Vehicle. IEEE Transactions on Vehicular Technology, Volume 61, 2012

[4] Xiong, R., Zhang, Y., Wang, J., He, H., Peng, S. and Pecht, M. Lithium-ion battery health prognosis based on a real battery management system used in electric vehicles. IEEE Transactions on Vehicular Technology, 2018

[5] Z. Lyu, R. Gao and L. Chen, "Li-Ion Battery State of Health Estimation and Remaining Useful Life Prediction Through a Model-Data-Fusion Method," in IEEE Transactions on Power Electronics, vol. 36, no. 6, 2021.

[6] B. Gou, Y. Xu, and X. Feng, "State-of-Health Estimation and Remaining-Useful-Life Prediction for Lithium-Ion Battery Using a Hybrid Data-Driven Method," in IEEE Transactions on Vehicular Technology, vol. 69, Oct. 2020.

[7] Hong, S., Hwang, H., Kim, D., Cui, S., Joe, I. "Real Driving Cycle-Based State of Charge 5. Prediction for EV Batteries Using Deep Learning Methods." Appl. Sci. 2021

[8] Y. Zhang, R. Xiong, H. He and M. G. Pecht, "Long Short-Term Memory Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-Ion Batteries," in IEEE Transactions on Vehicular Technology, vol. 67, July 2018.

[9] Jing Li, Yunjiao Wang, and Cheng Li, "Electric Vehicle State of Charge Estimation Based on Bidirectional Long Short-Term Memory Networks", IEEE Transactions on Industrial Electronics, Volume. 67, 2020.

[10] M. F. Niri, T. Q. Dinh, T. F. Yu, J. Marco and T. M. N. Bui, "State of Power Prediction for Lithium-Ion Batteries in Electric Vehicles via Wavelet-Markov Load Analysis," in IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 9, pp. 5833-5848, Sept. 2021.

[11] Y. Zhang, R. Xiong, H. He and W. Shen, "Lithium-Ion Battery Pack State of Charge and State of Energy Estimation Algorithms Using a Hardware-in-the-Loop Validation," in IEEE Transactions on Power Electronics, vol. 32, June 2017.

[12] C. Liu, K. Wang, Y. Zhang, and J. Qiu, "State-of-Charge Prediction for Electric Vehicles Based on a Recurrent Neural Network," in IEEE Access, vol. 8, pp. 182982-182991, 2020, doi: 10.1109/ACCESS.2020.3022533.

[13] Sanz-Gorrachategui, I., Pastor-Flores, P., Pajovic, M., Wang, Y., Orlik, P.V., Bernal-Ruiz, C., Bono-Nuez, A. and Artal-Sevil, J.S. "Remaining Useful Life Estimation for LFP Cells in Second-Life Applications". IEEE Transactions on Instrumentation and Measurement, 70, pp.1–10. doi:10.1109/tim.2021.3055791, 2021.

[14] Jun Xu; Chunting Chris Mi; Binggang Cao; Junjun Deng; Zheng Chen; Siqi Li, "The State of Charge Estimation of Lithium-Ion Batteries Based on a Proportional-Integral Observer", IEEE Transactions on Vehicular Technology (Vol no: 63, 2014)

[15] M. Xu, J. Peng, H. Sun and J. Xiong, "A Deep Recurrent Neural Network for EV State of Charge Prediction Based on Driving Context Data," in IEEE Transactions on Industrial Electronics, vol. 66,2019.

IJNRD Research Through Innovation