Enhanced Electric Vehicle Range Estimation Using TabNet and Advanced Machine Learning Techniques

Muhammed Shibil C V*, Amal MK†, and Rahul Satheesh‡
*Scool of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India

***Abstract*—This approach focuses on enhancing the range estimation of Electric Vehicles (EVs) by accurately predicting their energy consumption (EC) using advanced machine learning techniques. A unique aspect of our research is the application of TabNet, a state-of-the-art deep learning model specially developed for tabular data. In addition to this, we also explored a diverse set of traditional models, including LightGBM. We employed various advanced optimization techniques, such as Optuna and Bayesian optimization, to fine-tune these models for better performance. Furthermore, we developed ensemble models to leverage the strengths of individual algorithms and enhance overall forecasting accuracy. Our findings show the exceptional performance of TabNet in predicting EV energy consumption, demonstrating its potential as a powerful tool in this domain. LightGBM, when combined with clustering methods, also proved highly effective. The ensemble models we developed further improved prediction accuracy, offering robust and reliable estimates. This research represents a significant advancement in the field of sustainable transportation, providing a sophisticated approach to EV range estimation. By improving the accuracy of energy consumption predictions, our work addresses a crucial barrier to EV adoption, which is range anxiety, and supports the transition to cleaner, more efficient transportation systems.**

***Index Terms*—Electric Vehicles (EVs), Energy Consumption Rate (ECR), Machine Learning, TabNet, LightGBM, Bagging, Boosting, Neural Networks, Long Short-Term Memory (LSTM), Optuna, Bayesian Optimization, Ensemble Models, Sustainable Transportation.**

I. Introduction

About 25% of carbon dioxide emissions occur in the transport sector, making it one of the largest emitters of greenhouse gases [1]. The use of electric vehicles (EVs) is a widespread solution to reduce environmental pollution, and governments encourage their adoption over internal combustion engine cars [2]. EVs have gained significant attention as a sustainable solution for future transportation, driven by global concerns over climate change and international agreements like the Paris Agreement (2015) [3]. The popularity of EVs has increased competition among manufacturers, focusing on improving driving range capacity [4][5]. Global sales of EVs in 2018 increased by 72% compared to 2017, with a market share of 2.1%.

Despite these advantages, EV market share remains small due to high costs, long refueling times, and limited ranges [6]. Additional factors include insufficient battery technology [7] and a lack of a global network of charging stations, especially in developing countries [8]. A major challenge for EVs is accurately determining and increasing the driving range [9], which depends on the Energy Consumption Rate (ECR), expressed in kilowatt hours per hundred kilometers (kWh/100 km). That is, if an EV's ECR is 12 kWh per 100 km and the battery capacity is 30 kWh, it can cover a distance of 250 km [10].

The electric range (eRange) of an EV is crucial for consumers, providing an estimate of the remaining driving distance based on battery power and alleviating range anxiety [11][12]. The eRange can be estimated through various data parameters, including vehicle design, driver behavior, weather conditions, road inclination, and state of charge (SOC) estimation [13]. Accurate eRange estimates enable longer travel times and efficient charging plans. However, estimating eRange is complex due to multiple influencing factors.

Accurately predicting EV energy consumption is crucial to optimize their range, reduce range anxiety, and improve user acceptance. Traditional mathematical models often fail to capture the complex interactions affecting EV energy consumption, leading to inaccurate predictions. To address this, researchers have increasingly used machine learning techniques to learn from large datasets and make more accurate predictions. Machine learning, a subset of artificial intelligence, has shown success in various applications, including image recognition, natural language processing, and recommendation systems. By predicting EV energy consumption, machine learning can use data such as vehicle speed, battery charge status, ambient temperature, and traffic conditions to develop more accurate models. Algorithms like linear regression, decision trees, random forests, support vector machines, and artificial neural networks have been proposed for this purpose.

This work aims to select appropriate machine learning models to predict EV energy consumption by comparing each one with advanced optimization techniques and provide insight into the factors affecting their performance. The conclusions may facilitate the creation of more precise models and finding which optimization techniques can boost the performance of the model, aiding in the increased adoption of EVs and improving transportation sustainability. Machine learning's ability to learn from data and gradually improve results makes it an effective tool for solving complex problems [14][15][16][17]. This research investigates the application of machine learning techniques for accurate eRange estimation in EVs, contributing to the advancement of EV technology and promoting wider adoption of sustainable transportation solutions.

II. LITERATURE REVIEW

Recent studies have focused on improving energy consumption prediction for electric vehicles (EVs) to address range anxiety. Feng et al. [18] compared various regression models, finding that a Deep Multilayer Perceptron (MLP) achieved the best performance in estimating driving distance, with a mean absolute error of 5.58 km. Their Deep MLP classifier also showed 92.24% accuracy in predicting energy consumption rates. Pan et al. [19] proposed an innovative approach combining short-trip segment division with deep learning. By segmenting driving data into 10-second intervals and using Long Short-Term Memory (LSTM) networks, they captured complex relationships between driving conditions and energy consumption, accounting for real-world factors like traffic and driver behavior. Hosseini et al. [20] utilized big data and advanced machine learning techniques for mileage estimation. Their study, using a dataset of 35,000 real-world driving points, found that combinations of random forests with XGBoost and MLP, and Multiple Linear Regression with XGBoost, achieved high prediction accuracy.

Recent studies have explored different ways to predict the energy consumption of electric vehicles (EVs) to improve range estimation and promote sustainable transportation. Irfan Ullah et al. [21] compared a number of machine learning algorithms, including multi-linear regression, real-time root maximization, light intensification machine, and artificial neural networks. Their research used real-world driving data and common evaluation metrics to identify the most effective models for predicting electric vehicle energy consumption. Gurusamy et al. [22] reviewed a comprehensive analysis model for electric vehicle power components. His work highlighted the importance of accurately modeling transmission systems, electric motors, power converters and batteries. They highlighted the study by Zhang et al. (2019) and Miller et al. (2020) Li et al on engine efficiency. (2021), for the power loss converter, Wang et al. (2022) On the Modeling of Batteries. These studies highlight the need for integrated electronic models to improve energy consumption and range predictions.Amirkhani et al. [23] conducted a comparative study using real-time data from the Volkswagen e-Golf. They evaluated several different algorithms, including linear regression, multilayer perceptron, random forest, and adaptive propagation. Their research found the importance of considering factors such as average speed, road type, driving style and use of assistance systems. Similarly, Fetene et al. [24] use big data to estimate energy consumption and driving distance, and Wu et al. [25] investigated the influence of various factors on the energy expenditure rate.

The study discussed advanced methods for predicting electric vehicle (EV) energy consumption, highlighting the importance of hybrid machine learning approaches. Amirkhani et al. [23] used K-Fold cross-validation and cross-validation search grid for model optimization and evaluated the performance using metrics such as absolute error and R-squared score. Conditional importance analysis is an important insight into the design of hybrid models.The innovative framework presented in [25] addresses the key issues of electric vehicle driving and autonomy anxiety. This approach uses temporal sequential memory (LSTM) models and Transformer to efficiently capture time series data objects. The framework outperformed existing models by achieving an absolute percentage error (MAPE) of 4.63% in EV battery power prediction. In particular, when reflecting individual drivers and conditions, the MAPE decreased by 18.47% and 15.27%. In addition, this study proposed a strategy to achieve an MAPE of 6.7% by forecasting long-term electric vehicle energy consumption based on short-term data.Another important contribution [26] focuses on the prediction of short-term speeds and road gradients of hybrid electric vehicles (HEVs) using the autoregressive integrated moving average (ARIMA) model. This data-driven approach aims to improve the predictive energy management of HEVs. By incorporating ARIMA forecasts into energy management strategies, the study found a significant 5-7% reduction in HEV fuel consumption compared to a scenario without forecasts.

The research [27] focuses on improving resource efficiency in automobile traffic through modern driver assistance systems that minimize the vehicle's energy requirement through speed optimization algorithms. These systems rely on predictive route data to determine the energy requirements or upcoming operating points. It includes a method for predicting the energy demand of a hybrid electric vehicle using various data-based approaches, including neural networks, Gaussian processes and lookup tables. These approaches [27] are evaluated for their ability to predict the behaviour of individual powertrain components. The methods are trained using data from a test vehicle and the optimal approach is selected for each powertrain part is to provide the best prediction of energy demand. Validation results show that the method predicts the gear ratio of the transmission with an RMSE of 0.426, the torque of internal combustion engine with an RMSE of 19.01 Nm and the torque of the electric motor with an RMSE of 19.11 Nm. Furthermore, the root mean square error for motor voltage prediction is 1.211 V. This method [27] demonstrates the potential of data-driven approaches to optimize energy demand predictions, thereby contributing to more efficient operation of hybrid vehicles.

Another related work [28] addresses the challenge of accurately predicting energy consumption for new electric vehicle (EV) models using machine learning techniques. The authors developed a novel recommendation system to assist drivers on highways by predicting range of multiple electric vehicle models. Their approach leverages data-driven models trained on actual EV trip-driving data, achieving high accuracy for common EV models. However, due to limited trip-driving data for new EV models, the accuracy of predictions is lower. To solve this problem, the authors propose a transfer learning method that adapts prediction models of popular electric vehicles to new models, significantly reducing prediction errors by about 30%. This method promises to improve prediction accuracy for new EV models, which is essential for effective charging station recommendations and improving EV adoption on highways [28].

Introduced an ensemble stacked generalization (ESG) approach [29] that aims to improve the prediction accuracy of energy consumption of electric vehicles (EVs). ESG combines

multiple base regression models into a single meta regressor the strengths of individual algorithms– Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbor (KNN) to reduce model variance and improve predictive performance. By integrating data from Aichi Prefecture, Japan, including digital elevation maps and long-term GPS tracking data, the study estimates the energy efficiency (kWh/km) of electric vehicles based on critical factors such as average driving speed, distance, night lighting, air conditioning which includes Air conditioning consumption, heating ratio and road gradient. Evaluation using various statistical metrics confirms that ESG outperforms standalone models and demonstrates its robustness in predicting electric vehicle's energy consumption. These results [29] highlight the effectiveness of stacking techniques in improving prediction accuracy and provide valuable insights for urban infrastructure planning and the optimal use of charging facilities of electric vehicles.

Pokharel, S., et al. (2021) addresses the urgent need to accurately predict energy consumption in electric vehicles (EVs), which is crucial for alleviating range anxiety and and optimizing the deployment of charging infrastructure [30]. Using machine learning (ML) models such as multiple linear regression (MLR), extreme gradient boosting (XGBoost), and support vector regression (SVR), the study investigates total energy consumption (TEC) of electric vehicles in real world and external scenarios. Key independent variables include distance travelled, tire type, driving style, power, mileage, electric vehicle model and environmental factors such as city, highway and country roads. Among these, the distance travelled proves to be very influential and correlates strongly with the TEC (r=0.87). In particular, XGBoost achieves superior prediction accuracy (R^2=0.92), highlighting its effectiveness in addressing complex energy consumption patterns of electric vehicles. The results highlight the importance of robust predictive models for improving the usability of electric vehicles and guiding infrastructure development for sustainable urban mobility solutions [30].

Nabi, M. N., et al. (2023) work discussed the evolving automotive technology landscape amidst stringent emissions regulations aimed at mitigating the effects of global warming [31]. Hybrid electric vehicles (HEVs) have emerged as a provisional solution, combining combustion engines with electric motors powered by lithium batteries to improve fuel efficiency and reduce emissions. However, with increasing environmental concerns, there is an increasing shift towards electric vehicles (EVs) [31]. In this context, the study presents a comprehensive parametric analysis using a one – dimensional model developed for electric vehicles over different driving cycles using GT suite software. Important parameters such as engine power, battery charge level, vehicle speed, distance travelled and energy consumption are carefully evaluated. Additionally using neural network-based machine learning techniques, the paper predicts electric vehicle battery energy consumption with 89% accuracy, facilitating effective planning for electric vehicles drivers and enabling automotive engineers to innovate for more efficient and scalable electric vehicle designs [31].

Another important work of Zhu, Q., et al. (2024) introduced a novel machine learning-based energy consumption prediction framework for electric vehicles (EVs) that addresses the critical challenge of improving the accuracy to mitigate range anxiety and optimize energy consumption [32]. The research highlights the limitations of existing empirical, physics based and data-driven models and proposes an advanced approach that incorporates physics-informed functions and combines global offline models with vehicle-specific online customization [32]. By extensively testing this framework on real electric vehicle fleet data, the study highlights the superiority of the Quantile Regression Neural Network (QRNN) as a global model, achieving an average error reduction of 5.04% through online adjustment. In addition, the framework significantly improves prediction reliability, as evidenced by a 95% prediction interval coverage probability of 91.27% and an average prediction interval width reduced to 0.51. These results demonstrate the effectiveness of the framework in improving prediction accuracy and reliability and providing advances over existing methods [32].

This paper of Cabani, A., et al. (2021) explores a comprehensive approach to predicting the energy consumption rate of electric vehicles (EVs) by considering three main classes of influencing actors: environment, driver behaviour and vehicle parameters [33]. A novel model incorporating these classes and their interactions is developed to improve the quality of Electric Vehicle consumption predictions. The model uses a machine learning approach, specifically the k-Nearest Neighbors (k-NN) algorithm, and follows a lazy learning paradigm to improve the estimation performance [33]. This method benefits from using historical data to take the real ecosystem into account and capture factors such as driving style, heating, air conditioning use and battery health that directly impact Electric Vehicle energy consumption. The results show a high accuracy of 96.5% in estimating energy consumption and demonstrate the effectiveness of the proposed method. In addition to that, the model is used to optimize the energy efficient path between two points, providing a practical application for optimizing energy consumption [33].

The challenges hindering the widespread adoption of electric vehicles (EVs) Hua, Y., et al. (2022), particularly the lack of charging infrastructure and limited driving range, while emphasizing the importance of accurately estimating energy consumption. The proposed method aims to improve energy consumption predictions despite the challenges of insufficient EV data and irregular driving routes. It features a novel approach that includes knowledge transfer from Internal combustion engine/hybrid electric vehicles (ICE/HEV) to electric vehicles, segmentation assisted trajectory granularity, and time series estimation using a bidirectional recurrent neural network. Experimental evaluations show that this method outperforms other machine learning benchmarks in estimating energy consumption using a real-world electric vehicle energy dataset . The improved estimation can support better deployment of charging stations, improves greener driving behaviour and extend the range of electric vehicles, contributing to their wider adoption and environmental benefits.

III. Methodology

The method used in this study consists of several steps, including data processing, model selection, parameter optimization, model evaluation, and ensemble model construction. fig.5 represents the methodology diagram. The framework proposed for this study has several stages, starting with the data collection phase and providing an overview of the data set under investigation. This step is followed by exploratory data analysis (EDA), where issues such as missing values, outlier removal, image bias reduction, and key component selection are addressed. In this step, the data is thoroughly analysed using graphs and statistical methods to gain insights and prepare for the next steps. The model implementation phase involves building and evaluating several machine learning models. Started with simple linear regression model, That can be mathematically represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{1}$$

Where y is the predicted energy consumption, $x_1, x_2, \ldots, x_n$ are the features (e.g., speed, battery charge, temperature), β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and ϵ is the error term.

We tested many models including TabNet, LightGBM, Bagging, Boosting, Neural Networks and Short-Term Memory (LSTM). Over-the-counter optimization was performed to achieve optimal performance for each model. In particular, Bayesian optimization was used to tune the LightGBM model to improve the prediction accuracy. The performance of each model was evaluated based on metrics such as mean absolute error (MAE), mean squared error (MSE), and R-squared (R^2), as well as training and prediction times. The final step is to use the two best models to create an overall model using the cluster architecture. This approach aims to improve the performance and robustness of overall predictions by combining the strengths of individual models. Through this complex process, our framework ensures the evaluation and optimization of machine learning models for accurate and reliable predictions of power system consumption and anomaly detection.

A. Data Collection

Data plays an important role in predictive modelling. To investigate variables related to driving range in real EV scenarios, we use a dataset of 20 attributes listed in Table I. This data point comes from a reliable source https://www.spritmonitor.de.

B. Initial Data Preprocessing

The datasets used in this study were loaded from a CSV file and pre-processed to handle missing values and scale parameters. Pre-processing steps include separating features and target variables, processing numeric and categorical features separately, and applying transformations to make the data suitable for modelling numeric conditions, a pipe is created containing an input and a scale. The imputer replaces the missing values with the mean of that characteristic, and the scale averages the characteristic so that the mean and unit variance are zero. For the components, a tube is made in the facility with a unique heat coder. An imputer replaces the missing values with "absent" continuous values, and a co-heat encoder converts the phase features into a binary vector signal. Other preprocessing steps were implemented for TabNet, including extracting spatial features from the fuel dates (year, month, day) and encoding the segment parameters using the LabelEncoder. To calculate feature importance, the regression and selection process was performed using LightGBM Regressor, and features greater than 0 were selected for further processing. Quantitative regression using principal component analysis (PCA) was used to explain 95% of the variance in the data. For dimensionality reduction to d dimensions:

$$X_{\text{reduced}} = XW_d \tag{2}$$

where W_d is the matrix of the first d eigenvectors of the covariance matrix of X.

TABLE I
Attributes Influencing the Energy Consumption

Attributes	Description
Manufacturer	EV's company or brand
Model	Specific make and model
Version	Variant or edition of the model
Power (kW)	Motor's max power output in kW
Fuel Date	Date when EV was charged
Trip Distance (km)	Distance traveled on a charge in km
Quantity (kWh)	Total energy consumed in kWh
Fuel Type	Type of energy used (electricity)
Tire Type	Winter, summer, or all-year tires
City	Driven in the city or not
Motorway	Driven on motorways or not
Country Roads	Driven on country roads or not
Driving Style	Normal, moderate, or fast
A/C	Use of air conditioning
Park Heating	Use of vehicle heating
Avg Speed (km/h)	Average driving speed in km/h
ECR Deviation	Difference between recorded and manufacturer announced energy consumption
Fuel Note	Additional notes on fueling or performance

TABLE II
Target Variable

Variable	Description
Consumption (kWh/100km)	The rate of energy consumption of EV in kWh per 100 kilometres. It represents energy efficiency.

C. Data Augmentation

Data augmentation is a crucial technique in data science, enhancing the diversity and quality of datasets to improve model performance. In our research, we employed several augmentation techniques to refine and expand our dataset effectively.

1) Outlier Removal using Z-score: To begin with, we addressed the issue of outliers within our dataset. Outliers can significantly skew the results of statistical analyses and machine learning models, leading to inaccuracies. We utilized the Z-score method for outlier detection and removal. The Z-score measures the number of standard deviations a data

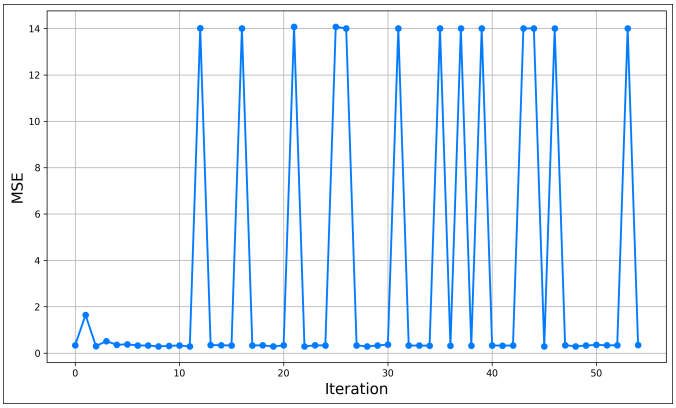

Fig. 1: Bayesian Optimization Convergence of LightGBM

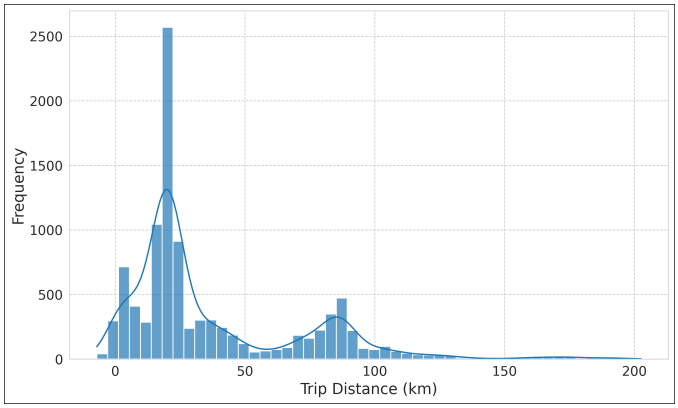

Fig. 3: Distribution of Trip Distances

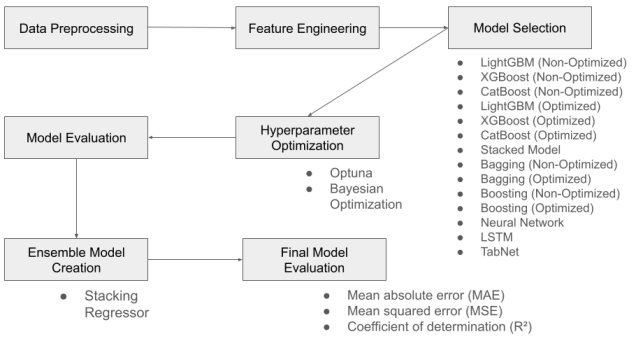

Fig. 2: Methodology diagram

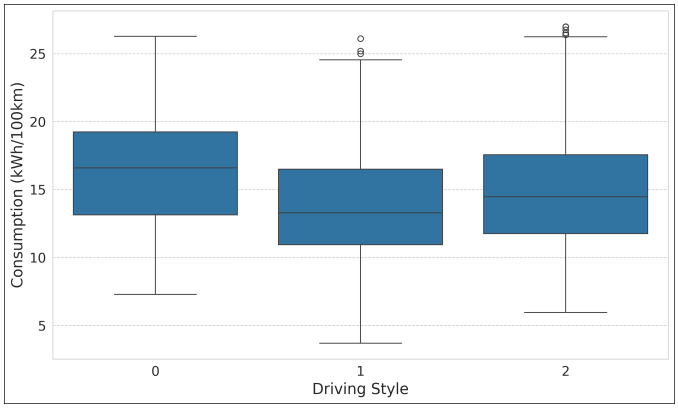

Fig. 4: Effect of Driving Style on Consumption

point is from the mean. Data points with Z-scores exceeding a threshold of 3 were identified as outliers and subsequently removed. This threshold ensures that only the most extreme outliers are excluded, retaining the integrity of the dataset while eliminating noise.

2) Synthetic Data Generation with Random Noise Addition: Following the outlier removal, we focused on generating synthetic data to augment our dataset. Synthetic data generation involved randomly sampling rows from the cleaned dataset and introducing random noise to numerical features. Specifically, we added random variations within predefined ranges to the features trip_distance(km), quantity(kWh), avg_speed(km/h), and consumption(kWh/100km). For instance, trip_distance(km) was perturbed by ±5 units, quantity(kWh) by ±2 units, avg_speed(km/h) by ±5 units, and consumption(kWh/100km) by ±1 unit. This method ensures the synthetic data points are similar to the original data but with slight variations, thereby increasing the dataset's diversity and robustness.

3) Combining Cleaned and Synthetic Data: The final step in our data augmentation process involved merging the cleaned dataset with the newly generated synthetic data. By concatenating these two datasets, we created an augmented dataset that combines the benefits of clean data with the added diversity of synthetic data points. This comprehensive dataset is better suited for training machine learning models, as it encompasses a broader range of variations while maintaining data quality.

D. Exploratory Data Analysis

Performed a comprehensive EDA including univariate, bivariate and multivariate analyses. This includes understanding data distributions and trends, identifying potential trends, and choosing the optimal number of informative features. This step is an important preparatory step for the following steps: technical development and prototype development. For multiple EDAs. Analyzed the relation between dependent variables and target as shown in Fig. 1 the distribution of trip distances with respect to Frequency. The Effect of driving style on consumption is shown in Fig. 2, Respectively Distribution of consumption is shown in Fig. 4.

E. Hyperparameter Optimization

In this study, we perform linear optimization using the following methods:

1) Optuna: Optuna is a linear optimization framework that uses tree-based Parzen estimation to find the best dollar combination. It is used to optimize TabNet, LightGBM, XGBoost and CatBoost models by defining an objective function that minimizes the worst case during cross-validation, Optuna aims to minimize an objective function $f(\theta)$, where θ is the set of hyperparameters. and it can be mathematically represented as:

$$\theta^* = \arg\min_{\theta \in \Theta} f(\theta) \tag{3}$$

Where θ are the hyperparameters, and $f(\theta)$ is the loss function to be minimized.

2) Bayesian Optimization: Bayesian optimization is a model-based optimization technique that uses probabilistic models to guide the search for the optimal set of parameters. It is used to optimize the LightGBM model by defining an objective function that minimizes the worst-case error during cross-validation. Bayesian optimization algorithms update a probability model based on the observed model function and use it to generate a new set of parameters to evaluate. and it can be mathematically represented as:

$$\theta^* = \arg\max_{\theta \in \Theta} E[f(\theta)] \tag{4}$$

Where θ represents the hyperparameters, and $f(\theta)$ is the objective function to be maximized.

These two-dimensional optimization techniques are used to improve the performance of the selected models, ensuring that the optimal combination is selected for each model.

F. Model Selection

In this study, several machine learning models are used for prediction tasks. Here is an explanation of each model mentioned in the three parts:

1) TabNet: We also implemented TabNet Regressor, a deep learning model for tabular data, using the pytorch_tabnet library. TabNet was chosen for its ability to handle mixed data types and its state-of-the-art performance on tabular data. The model was trained with a patience of 10 and a maximum of 100 epochs. Key TabNet hyperparameters include:

n_d : Number of decision steps
n_a : Number of attention steps
γ : Relaxation parameter
$n_{\text{independent}}$: Number of independent GLU layers
n_{shared} : Number of shared GLU layers
momentum : Momentum for batch normalization
mask_type : Type of mask used (e.g., entmax, sparsemax)

For a given input x:

$$\text{Decision steps: } d_t = \text{GLU}(d_{t-1}, x) \tag{5}$$

$$\text{Attention steps: } a_t = \text{GLU}(a_{t-1}, x) \tag{6}$$

$$\text{Output: } y = \sum_{t=1}^{T} \gamma_t d_t \tag{7}$$

where GLU denotes the Gated Linear Unit and T is the total number of steps.

2) LightGBM: It is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be efficient in terms of both memory and computational resources. Optuna is a hyperparameter optimization framework used to find the best set of hyperparameters for the LightGBM model. it can be can be mathematically represented as:

$$f(x) = \sum_{i=1}^{M} T_i(x; \theta_i) \tag{8}$$

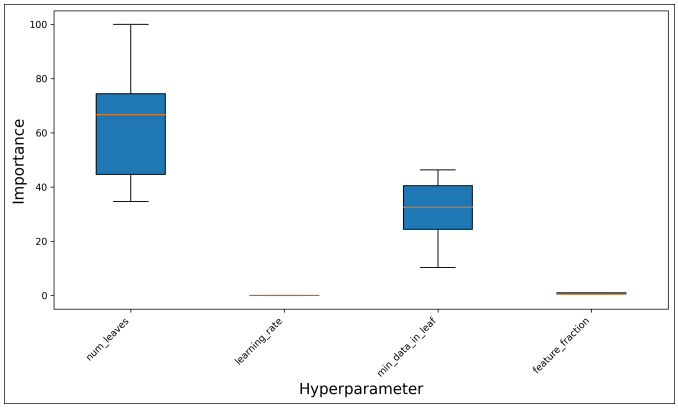

Fig. 5: importance of hyperparameters derived from the results of Bayesian optimization for LightGBM model

Where $f(x)$ is the prediction function, T_i represents individual decision trees, θ_i are the parameters of the trees, and M is the number of trees.

3) Bagging Regressor: Bagging Regressor is an ensemble learning method that combines the predictions of multiple base models to improve the overall performance. It trains several estimators in parallel and aggregates their predictions to reduce variance and prevent overfitting.

4) Gradient Boosting Regressor: Gradient Boosting Regressor is another ensemble learning method that builds multiple weak models sequentially, with each model attempting to correct the errors of the previous model. It works by minimizing the loss function using gradient descent.

5) Neural Network: A Neural Network is a powerful machine learning model that can learn complex relationships between features and target variables. It consists of interconnected layers of nodes (neurons) that process and transmit information to make predictions. And it can be mathematically represented as:

$$y = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{9}$$

Where y is the output, σ is the activation function, w_i are the weights, x_i are the inputs, and b is the bias term.

6) LSTM: Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that can learn and remember long-term dependencies in sequential data. LSTM networks are particularly useful for time series prediction tasks.

7) LightGBM (Optuna and Bayesian Optimization): In this case, both Optuna and Bayesian Optimization are used to find the best set of hyperparameters for the LightGBM model. Bayesian Optimization is a sequential model-based optimization technique that uses a probabilistic model to guide the search for the best set of hyperparameters.

These models are selected due to their effectiveness in handling heterogeneous data and optimizing predictive accuracy. The performance of each model is evaluated using appropriate metrics to ensure a comprehensive comparison and select the best model for the specific prediction task.

G. Ensemble Model Creation

In machine learning, an ensemble model is a combination of multiple individual models to improve overall performance. Ensemble models can be created using various techniques, such as bagging, boosting, and stacking. In the provided codes, the following ensemble models are created:

1) Bagging Regressor: Bagging (Bootstrap Aggregating) is an ensemble technique that involves creating multiple subsets of the original dataset, training a base model on each subset, and aggregating their predictions. Bagging Regressor is an implementation of this technique using decision trees as the base estimator. It reduces variance and prevents overfitting by introducing randomness in the data sampling and feature selection processes.

2) Gradient Boosting Regressor: Gradient Boosting is an ensemble technique that involves training multiple weak models sequentially, with each model attempting to correct the errors of the previous model. Gradient Boosting Regressor is an implementation of this technique using decision trees as the base estimator. It works by minimizing the loss function using gradient descent.

3) Stacking Ensemble: Stacking is an ensemble technique that involves training multiple base models on the same dataset and combining their predictions using a meta-model (also known as a second-level model). In the provided codes, a stacking ensemble is created using LightGBM, XGBoost, CatBoost, and a Neural Network as base models, and a GradientBoostingRegressor as the meta-model. The predictions of the base models are used as input features for the meta-model, which learns to combine them optimally.

Ensemble models can often achieve better performance than individual models by leveraging the strengths of each model and reducing the impact of their weaknesses. The choice of ensemble technique depends on the specific problem, dataset, and performance metrics. In the provided codes, ensemble models are created to improve the predictive accuracy of electric vehicle energy consumption.

This study employs a comprehensive methodology to preprocess the data, select and optimize the models, evaluate their performance, and create an ensemble model to improve the predictive accuracy of electric vehicle energy consumption. The results are expected to offer insights into optimal model selection and configuration, contributing to the advancement of predictive modelling for sustainable transportation

IV. Results and Discussion

In this study, we use a unique hybrid machine learning approach to calibrate electric vehicle (EV) energy consumption predictions. We use various techniques of data processing, model selection, parameter optimization, model evaluation, and ensemble model building. We thoroughly review several machine learning models, focusing on predictive accuracy and computational efficiency.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

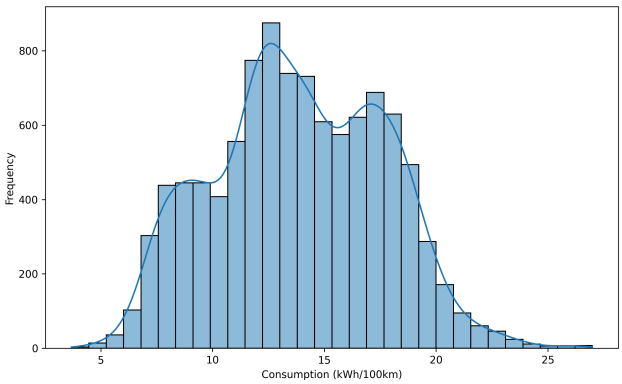

Fig. 6: Distribution of Consumption (kWh/100km)

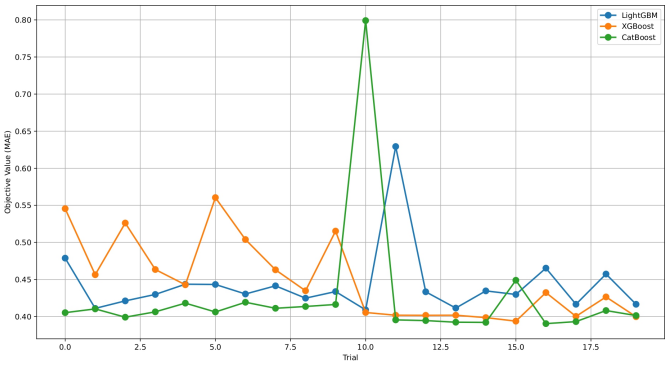

Fig. 7: Optimization History

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

We implemented and evaluated several machine learning models, including LightGBM, XGBoost, CatBoost, neural

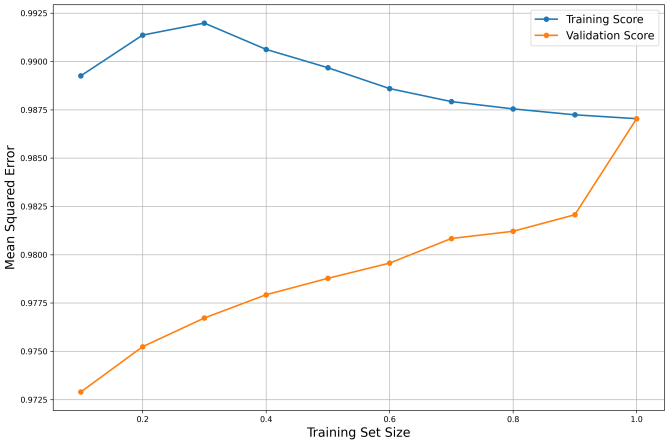

Fig. 8: Learning curves for a LightGBM regression model

TABLE III
Model Performance Comparison

Model	MAE	MSE	R^2	Training Time (s)	Prediction Time (s)
Optimized LightGBM (Bayesian)	0.387923	0.274267	0.980780	0.475633	0.027214
Optimized LightGBM (Optuna)	0.389102	0.277539	0.980551	0.258062	0.020459
Optimized Boosting (Optuna)	0.385846	0.278497	0.980484	1.656084	0.005594
Bagging	0.387068	0.277644	0.980544	1.710479	0.154376
Boosting	0.387231	0.279364	0.980423	1.466153	0.005745
Non-Optimized LightGBM	0.392147	0.281602	0.980266	0.340715	0.0438917
XGBoost (Non-Optimized)	0.391060	0.282114	0.980230	0.393101	0.0368435
CatBoost (Non-Optimized)	0.396548	0.281007	0.980308	7.221510	0.0333936
Stacked Model	0.394417	0.280670	0.980332	22.8626	0.0625374
Optimized XGBoost (Optuna Optimized)	0.394479	0.280121	0.980370	0.491883	0.0299349
CatBoost (Optuna Optimized)	0.394494	0.283894	0.980106	2.275930	0.0595214
LightGBM	0.393093	0.282047	0.980235	0.235040	0.017244
LSTM	0.423800	0.310386	0.978249	5.174473	0.755519
Optimized Bagging (Optuna)	0.436971	0.342638	0.975989	4.013618	0.070040
TabNet	0.495946	0.407326	0.971456	5.912781	NIL
Optimized TabNet	0.502809	0.420081	0.970562	5.197525	NIL
Neural Network	1.023315	1.300441	0.908869	0.704010	0.248364

networks, and short-term memory (LSTM) networks. Multidomain optimization was performed using the Optuna and Bayesian optimization methods. Optuna, which uses a tree-based Parzen estimator, was used to optimize the LightGBM, XGBoost, and CatBoost models. We further improved the LightGBM model using Bayesian optimization, a model building technique that uses probabilistic models.The performance of each model was evaluated using absolute error (MAE), mean squared error (MSE) and coefficient of determination (R2). We also recorded the training time and the prediction time to evaluate the efficiency of the software.Among the optimized models, the LightGBM model optimized using Bayesian optimization showed the best performance. The model achieved an impressive R2 score of 0.980780, indicating a high level of accuracy in capturing the underlying patterns in the data. The MAE is 0.387923 and the MSE is 0.274267, which shows the ability to reduce the prediction error. In addition, the optimized LightGBM model showed strong computational performance with a training time of 0.475633 seconds and a prediction time of 0.027214 seconds.The LightGBM model optimized using Optuna performed very well, with an R2 score of 0.980551, MAE 0.389102, and MSE 0.277539. This model has a faster training time of 0.258062 seconds and a faster prediction time of 0.020459 seconds, making it a good choice for real-time applications.We also review good and bad promotional examples. Optuna's optimization model achieved an R2 score of 0.980484, MAE of 0.385846, and MSE of 0.278497. The training time is 1.656084 seconds, but the prediction time is as short as 0.005594 seconds, which is suitable for situations that require fast prediction.Although the jump model did not perform well in terms of accuracy, it showed competitive results with an R2 score of 0.980544, MAE 0.387068, and MSE 0.277644. The training time is 1.710479 seconds and the prediction time is 0.154376 seconds, maintaining a balance between accuracy and computational efficiency.We also reviewed several popular models for gradient boosting, such as LightGBM, XGBoost, and CatBoost. The ensemble model, which combined several algorithms to improve prediction accuracy, achieved an R2 score of 0.980332 and an MAE of 0.394417. Although computationally more expensive, clustering techniques have proven to be powerful in capturing complex relationships in data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

We also looked at other models such as LSTM and neural networks. The LSTM model using TensorFlow showed good results and the neural network model implemented using Keras showed accuracy despite the long training and prediction time. The analysis of both optimizations are shows that performance of the models gradually increased by optimiser in the (fig. 1) shows the Bayesian optimizations convergence of LightGBM and in (fig. 4) shows the importance of hyperparameters derived from the results of Bayesian optimization for LightGBM model. In (fig. 7) the optimization history of 3 models are plotted respectively. also in addition to that Learning curves for LigtGBM regression model is also plotted in (fig. 8). In conclusion, our study demonstrated that a hybrid machine learning approach can improve energy consumption prediction. Model optimization using techniques such as Optuna and Bayesian optimization improved prediction accuracy while accounting for computational efficiency. These results highlight the importance of comparing performance metrics and considering their implementation in data-driven decision-making processes. Our comprehensive framework ensures accurate and reliable forecasts of electric vehicle energy consumption, contributing to the advancement of sustainable transport.

V. Conclusion

In this study, we aim to refine electric vehicle (EV) energy consumption predictions using a unique combination of machine learning techniques. Our primary goal is to improve the accuracy and efficiency of prediction models, to pave the way for more sustainable energy efforts in the energy system. We tested a wide range of machine learning models, from gradient

boosting methods like LightGBM and XGBoost to clustering methods like Bagging and Stacking. We further improved these models by optimizing several parameters using Optuna and Bayesian methods to improve performance metrics, including absolute error (MAE), mean squared error (MSE), and coefficient of determination (R2).Among the models evaluated, the LightGBM model optimized by Bayesian optimization stood out with an impressive R2 score of 0.980780. This shows its great ability to capture complex relationships in electric vehicle energy consumption data. This model shows good computational efficiency and other variations, making it suitable for real-time applications in electronic systems..In conclusion, optimizing the prediction of electric vehicle energy consumption using a hybrid machine learning approach is an important step forward to achieve a sustainable and efficient electric power system. By harnessing the power of data-driven insights, we aim to create a future in which electric vehicles play an important role in driving sustainability and environmental innovation.

REFERENCES

[1] Y. Huang, H. Wang, A. Khajepour, H. He, and J. Ji, "Model predictive control power management strategies for HEVs: A review," *Journal of Power Sources*, Elsevier B.V., Feb. 2017. Available: https://doi.org/10.1016/j.jpowsour.2016.11.106

[2] M. Bohlsen, "EV Company News for the Month of January 2019," *Seeking Alpha*, 2019.

[3] M. Figenbaum, A. Gopal, and M. Sivak, "Market competition and technology diffusion: The case of electric vehicles," *Transportation Research Part D: Transport and Environment*, vol. 37, pp. 142–153, 2015.

[4] C. O. Egbue and S. Long, "An analysis of factors influencing electric vehicle adoption in the United States," *Transportation Research Part D: Transport and Environment*, vol. 17, no. 1, pp. 38–45, 2012.

[5] A. Amirkhani, A. Haghanifar, and M. R. Mosavi, "Electric Vehicles Driving Range and Energy Consumption Investigation: A Comparative Study of Machine Learning Techniques," in *Proc. 5th Iranian Conf. Signal Processing and Intelligent Systems, ICSPIS*, 2019. Available: https://doi.org/10.1109/ICSPIS48872.2019.9066042

[6] J. Hong, S. Park, and N. Chang, "Accurate remaining range estimation for Electric vehicles," in *Proc. Asia and South Pacific Design Automation Conf., ASP-DAC*, Jan. 2016, pp. 781–786. Available: https://doi.org/10.1109/ASPDAC.2016.7428106

[7] J. Axsen, K. S. Kurani, and A. Burke, "Are batteries ready for plugin hybrid buyers?," *Transp. Policy*, vol. 17, no. 3, pp. 173–182, May 2010.

[8] E. Narassimhan and C. Johnson, "The role of demand-side incentives and charging infrastructure on plug-in electric vehicle adoption: Analysis of US States," *Environmental Research Letters*, vol. 13, no. 7, 2018. Available: https://doi.org/10.1088/1748-9326/aad0f8

[9] G. Wager, J. Whale, and T. Braunl, "Driving electric vehicles at highway speeds: The effect of higher driving speeds on energy consumption and driving range for electric vehicles in Australia," *Renewable and Sustainable Energy Reviews*, Elsevier Ltd., Sept. 2016. Available: https://doi.org/10.1016/j.rser.2016.05.060

[10] G. M. Fetene, S. Kaplan, S. L. Mabit, A. F. Jensen, and C. G. Prato, "Harnessing big data for estimating the energy consumption and driving range of electric vehicles," *Transportation Research Part D: Transport and Environment*, vol. 54, pp. 1–11, Apr. 2017. Available: https://doi.org/10.1016/j.trd.2017.04.013

[11] Y. Song and Y. Hu, "Range Anxiety of Electric Vehicle Considering Charging Infrastructure and Battery Degradation," *IEEE Trans. on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3108–3118, 2021.

[12] M. Varga, G. Bálint, and P. Madarász, "Comparison of electric vehicle range estimation methods," *Periodica Polytechnica Transportation Engineering*, vol. 47, no. 1, pp. 41–48, 2019.

[13] T. Condie, E. Curry, and N. Morgan, "MapReduce for machine learning on multicore," *Journal of Parallel and Distributed Computing*, vol. 73, no. 3, pp. 421–431, 2013.

[14] Z. Zhou, P. Wang, Q. Zhang, and B. Li, "A survey of big data machine learning for smart grids," *IEEE Trans. on Smart Grid*, vol. 8, no. 6, pp. 2824–2837, 2017.

[15] I. Bose and S. S. Mahapatra, "A review of some recent developments in data mining," *Sadhana*, vol. 26, no. 1, pp. 41–74, 2001.

[16] T. M. Mitchell, *Machine Learning*, McGraw Hill, 2006.

[17] Y. Pan, W. Fang, and W. Zhang, "Development of an energy consumption prediction model for battery electric vehicles in real-world driving: A combined approach of short-trip segment division and deep learning," *Journal of Cleaner Production*, vol. 400, 2023. Available: https://doi.org/10.1016/j.jclepro.2023.136742

[18] K. Unni and S. Thale, "Energy Consumption Analysis for the Prediction of Battery Residual Energy in Electric Vehicles," *Engineering, Technology and Applied Science Research*, vol. 13, no. 3, pp. 11011–11019, 2023. Available: https://doi.org/10.48084/etasr.5868

[19] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal, "A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability," *Energy and Environment*, vol. 33, no. 8, pp. 1583–1612, 2022. Available: https://doi.org/10.1177/0958305X211044998

[20] A. Gurusamy, B. Ashok, and B. Mason, "Prediction of Electric Vehicle Driving Range and Performance Characteristics: A Review on Analytical Modeling Strategies With Its Influential Factors and Improvisation Techniques," *IEEE Access*, vol. 11, pp. 131521–131548, 2023. Available: https://doi.org/10.1109/ACCESS.2023.3334620

[21] Z. Feng, J. Zhang, H. Jiang, X. Yao, Y. Qian, and H. Zhang, "Energy consumption prediction strategy for electric vehicle based on LSTM-transformer framework," *Energy*, vol. 302, 2024. Available: https://doi.org/10.1016/j.energy.2024.131780

[22] X. Wu, D. Freese, A. Cabrera, and W. A. Kitch, "Electric vehicles' energy consumption measurement and estimation," *Transportation Research Part D: Transport and Environment*, vol. 34, pp. 52–67, 2015. Available: https://doi.org/10.1016/j.trd.2014.10.007

[23] S. Hosseini, A. Yassine, and T. Akilan, "Ensemble-based Robust Model for Accurate Driving Range Estimation of EVs Leveraging Big Data," 2024. Available: https://doi.org/10.1109/ENERGYCON58629.2024.10488792

[24] M. C. Smuts, P. Bertoldi, and W. C. Knottenbelt, "Range anxiety in electric vehicles: A review of factors influencing range and strategies to mitigate range anxiety," *Applied Energy*, vol. 204, pp. 887–901, 2017.

[25] J. Guo, H. He, and C. Sun, "ARIMA-based road gradient and vehicle velocity prediction for hybrid electric vehicle energy management," *IEEE Trans. on Vehicular Technology*, vol. 68, no. 6, pp. 5309–5320, 2019. Available: https://doi.org/10.1109/TVT.2019.2912893

[26] D. Fink, O. Maas, D. Herda, Z. Ziaukas, C. Schweers, A. Trabelsi, and H. G. Jacob, "Data-Based Energy Demand Prediction for Hybrid Electrical Vehicles," *SN Computer Science*, vol. 5, no. 1, 2024. Available: https://doi.org/10.1007/s42979-023-02475-9

[27] A. Fukushima, T. Yano, S. Imahara, H. Aisu, Y. Shimokawa, and Y. Shibata, "Prediction of energy consumption for new electric vehicle models by machine learning," *IET Intelligent Transport Systems*, vol. 12, no. 9, pp. 1174–1180, 2018. Available: https://doi.org/10.1049/iet-its.2018.5169

[28] I. Ullah, K. Liu, T. Yamamoto, M. Zahid, and A. Jamal, "Electric vehicle energy consumption prediction using stacked generalization: an ensemble learning approach," *International Journal of Green Energy*, vol. 18, no. 9, pp. 896–909, 2021. Available: https://doi.org/10.1080/15435075.2021.1881902

[29] S. Pokharel, P. Sah, and D. Ganta, "Improved prediction of total energy consumption and feature analysis in electric vehicles using machine learning and shapley additive explanations method," *World Electric Vehicle Journal*, vol. 12, no. 3, 2021. Available: https://doi.org/10.3390/wevj12030094

[30] M. N. Nabi, B. Ray, F. Rashid, W. Al Hussam, and S. M. Muyeen, "Parametric analysis and prediction of energy consumption of electric vehicles using machine learning," *Journal of Energy Storage*, vol. 72, 2023. Available: https://doi.org/10.1016/j.est.2023.108226

[31] Q. Zhu, Y. Huang, C. Lee, et al, "Predicting Electric Vehicle Energy Consumption from Field Data Using Machine Learning," *IEEE Trans. on Transportation Electrification*, 2024. Available: http://dx.doi.org/10.1109/TTE.2024.3416532

[32] A. Cabani, P. Zhang, R. Khemmar, and J. Xu, "Enhancement of energy consumption estimation for electric vehicles by using machine learning," *IAES Int. J. Artificial Intelligence*, vol. 10, no. 1, pp. 215–223, 2021. Available: https://doi.org/10.11591/ijai.v10.i1.pp215-223