Carbon Circularity
https://ojs.ukscip.com/index.php/cc

ARTICLE

Household-Level Carbon Footprint Forecasting in Nigeria: A Machine Learning Approach with Prediction Error Risk Assessment for Net-Zero Emissions

Lanre Olatomiwa [1*], *Harrison Oyibo Idakwo* [1,2], *James Garba Ambafi* [1], *Umar Suleiman Dauda* [1], *Isiyaku Saleh* [1], *Kufre Esenowo Jack* [3]

[1] *Department of Electrical & Electronics Engineering, Federal University of Technology Minna, Minna 920101, Nigeria*
[2] *Department of Electrical & Electronics Engineering, University of Maiduguri, Maiduguri 600104, Nigeria*
[3] *Department of Mechatronics Engineering, Federal University of Technology Minna, Minna 920101, Nigeria*

ABSTRACT

The study develops a data-driven framework for predicting household CO_2 emissions within a developing economic setting using Talba Estate in Minna, Niger State, Nigeria, as a case study. Hourly data were collected from 10 households for the whole of 2023, encapsulating electricity consumption, income, household size, and climatic parameters. Four machine learning models were benchmarked and evaluated within a prediction-uncertainty risk assessment framework, which quantifies the likelihood and impact of model-based prediction errors rather than policy or environmental risks. The models were trained on a 70/30 train-test split and evaluated within a novel prediction-error risk assessment framework that quantifies model uncertainty. The XGBoost achieved the highest in predictive accuracy among the four, with minimum error rates: MAPE = 0.0073, RMSE = 0.1463, and MAE = 0.0340, an R^2 of 0.9999, almost a perfect fit. The robustness of the model was also tested by prediction-error risk scoring, with values averaging around zero and stability values at about 0.100 across households. The key innovation is the integration of machine learning forecasting with a structured prediction-error risk assessment frame-

*CORRESPONDING AUTHOR:
Lanre Olatomiwa, Department of Electrical & Electronics Engineering, Federal University of Technology Minna, Minna 920101, Nigeria; Email: olatomiwa.l@futminna.edu.ng

ARTICLE INFO
Received: 5 November 2025 | Revised: 17 December 2025 | Accepted: 19 December 2025 | Published Online: 20 December 2025
DOI: https://doi.org/10.54963/cc.v1i1.1861

CITATION
Olatomiwa, L., Idakwo, H.O., Ambafi, J.G., et al., 2025. Household-Level Carbon Footprint Forecasting in Nigeria: A Machine Learning Approach with Prediction Error Risk Assesment for Net-Zero Emissions. Carbon Circularity. 1(1): 25–42.
DOI: https://doi.org/10.54963/cc.v1i1.1861

work, applied to high-resolution household data in a resource-constrained setting, a combination rarely addressed in existing literature. The results point toward a promising outlook for hybridizing an advanced machine-learning toolkit with prediction-uncertainty risk quantification toward accurate carbon forecasting in resource constraint context. The findings offer actionable insights for policymakers supporting Sustainable Development Goal 13 and Nigeria's net-zero emissions targets, advancing scalable carbon monitoring frameworks for developing regions.
Keywords: Machine Learning; Carbon Footprint; Prediction; Emission; Prediction-Error Risk Assessment; Model Uncertainty; Nigeria

1. Introduction

Currently, the urgency of fighting climate change has put forth carbon footprints measuring and reducing as a good area for sustainability-related research wherein carbon footprint has been a common key parameter to gauge greenhouse gas emissions and consequently monitor the progress line of Sustainable Development Goals (SDGs), especially Goal 13 of the SDGs: Climate Action [1–3]. The emission reduction issue gained relevance after the conclusion of the Paris Climate Agreement, which puts forward that global warming should be kept below 2 ℃ above pre-industrial levels and, if possible, brought down to 1.5 ℃ [4–6]. The most significant types of greenhouse gas (GHG) emissions, that is carbon dioxide consitute 81 per cent, therefore calling for proper monitoring and studying in the interest of environmental policy and climate action [7,8].

An estimate of carbon emissions helps policymakers to set realistic targets for carbon reduction, to assess the efficiency of the strategies being pursued, and to draw balanced and rational policies that may promote economic growth and environmental protection subject to the given geopolitical scenario [9]. By explanation of the system mechanism, the carbon emissions are complicated and result from many factors, including population growth, economic growth, energy use, technology changes, and alterations to energy systems. By research, without any solution, current levels of emissions will lead to the worst-case scenarios, and environmental degradation is severe, such as more frequent and severe floods, fires, and storms by 2100 [10].

The demand to amplify prediction skill and increased availability of data have resulted, over the years, in the enhancement of a variety of prediction models, from simple statistics to more sophisticated AI techniques [11]. Secondly, when considering the regional and global emission trends, some strategies need to be contemplated for forecasting and limiting carbon emissions. Developed and developing economies follow different emission trajectories since those in the developing have their emissions tend to grow with processes of industrialization and urbanization. This difference might reveal the contextual variables useful for the designing of models in line with the socioeconomic contexts. Besides, even after the very rich literature had immensely contributed towards retrospective carbon accounting, life cycle assessment, and sectoral drivers of emissions, these approaches are still mostly descriptive and, therefore, cannot estimate future emission scenarios [12,13].

Addressing this forecasting gap is complex, as it exists within an uncertain socioeconomic landscape where technological advancement and policy implementation are variable. This uncertainty itself presents a major impediment to formulating robust risk mitigation strategies aligned with national and international net-zero pathways. Although machine learning (ML) has revolutionized emission forecasting at corporate, industrial, and national scales [11], its application to critical residential sector, particularly in developing economies with distinct socioeconomic pressures, faces notable limitations.

Significant and interconnected gaps persist in the extant literature. Firstly, while ML applications are growing, a disproportionate focus remains on corporate, industrial and national level emissions [14–16], with granular household-level forecasting in developing economies remaining underexplored. Secondly, much of the present literature is context-specific, aimed at developed economies with stable data infrastructures, while

developing contexts like Nigeria are largely overlooked despite rising emissions. Thirdly and most critically, existing studies predominantly prioritize predictive accuracy but seldom integrate forecasting outputs into a formal, decision centric risk management framework. Few models account for the systemic risk arising from socioeconomic, climatic, and policy-induced uncertainties, limiting their practical utility for policy and planning.

Addressing these gaps, this study introduces an integrated framework that advances the literature in three key ways: i.) employing a variety of advanced machine learning algorithms (RFR, ETR, EVR and XGBoost) for comparative benchmarking; ii.) Integrating predictions into a structured prediction-error risk assessment model to quantify uncertainties; and iii.) situating this analysis within a developing economy context, where balancing socioeconomic growth with carbon footprint reduction is paramount.

The major innovation in this study is in its integration of machine-learning-based carbon-emission prediction with a structured risk-management framework, using high-resolution real time household CO_2 data from a developing–country context (Nigeria), benchmarking multiple advanced models, and providing risk-linked mitigation strategies that support SDG 13 and net-zero decision-making.

The primary aim of this study is to develop a robust, risk-informed framework for predicting household CO_2 emissions in a resource-constrained setting. This aim is achieved through the following objectives:

- Develop machine-learning models (RFR, ETR, SVR and XGBoost) capable of accurately predicting household CO_2 emission using environmental, socioeconomic and energy-consumption variables.
- Integrate a prediction-error risk assessment model into the prediction framework in order to quantify uncertainties in household CO_2 emission forecasts.
- Evaluate and compare the predictive performance of the selected machine learning models across households in a developing-economy context.
- Determine the likelihood, impact and overall risk score associated with prediction errors for each household under the proposed prediction-error risk assessment framework.
- Recommend mitigation strategies and actionable policy pathways for reducing household-level carbon emission while supporting SDG13 and net- zero aspirations.

And the corresponding research questions derived directly from the objectives are:

1. How accurately can machine-learning models predict household CO_2 emissions using the selected environmental, socioeconomic, and energy-consumption variables?
2. How can prediction-error risk assessment principles be integrated into carbon- emission prediction models to quantify uncertainty in household forecasts?
3. Which of the four models (RFR, ETR, SVR, XGBoost) demonstrates superior accuracy and robustness when applied to household-level CO_2 datasets in a developing country environment?
4. What are the likelihood, impact and resulting risk levels associated with prediction errors for each household under the prediction-error risk assessment framework?
5. What mitigation measures and policy recommendations can be derived from the prediction and prediction-error risk assessment results to support climate-action and net-zero targets?

The rest of this paper is structured as follows: Section 2 presents the literature review and the related works. Section 3 deals with the methodology and discusses the forecasting approaches and the framework for assessing risks. Section 4 shows the results and the interpretation of emission patterns and related risks. Section 5, therefore, not only concludes the discussion by presenting some main insights and wider implications, but also, discusses limitations and possible avenues for future work.

2. Literature Review

2.1. Traditional and Conventional Approaches to Carbon Accounting

The foundational work in carbon footprint analysis has largely relied on descriptive and retrospective ap-

proaches. Life Cycle Assessment (LCA) and retrospective carbon accounting have provided critical insights into sectoral contributions to emissions and have been instrumental in establishing baseline inventories[13]. However, as noted by Li et al.[12], these approaches are inherently descriptive and are limited in their capacity to forecast future emission scenarios under uncertainty. They provide a vital historical picture but offer limited utility for proactive, forward looking policy formulation and risk mitigation.

2.2. The Rise of Data-Driven and Machine Learning Forecasting Models

To address the need for prediction, the field has increasingly turned to data-driven and machine learning (ML) techniques. Recent years have seen the successful application of ML models for forecasting emissions at various scales. Nguyen et al.[15] demonstrated the use of models like Elastic Net, Random Forest, and XGBoost to forecast corporate carbon emissions, though they noted the limitation of not extending to household footprints. Similarly, studies have applied AI for optimizing supply chain emissions[16] and creating high-precision models for industrial sources like coal-fired power plants[17]. At a macro-scale, advanced time-series approaches like SARIMA have been used to model global emissions, incorporating disruptions such as the COVID-19 pandemic[18]. This demonstrates a clear trend towards using sophisticated algorithms to model the complex, non-linear drivers of emissions across different domains.

2.3. Review of Related Work

Nguyen et al.[15] considered the use of machine learning models in forecasting carbon emissions with a large set of predictors—attributes of the firm, industrial types, and several environmental elements. The model, which was built with some degree of thoroughness with predictors such as Elastic Net, Random Forest, and XGB models, even underwent out-of-sample validation, enhancing the trustworthiness of the study. Industrial activity and fuel mix were also considered as emissions factors largely in the study. The authors, however, denote that analysis was limited to corporate and industry emissions and did not extend to household carbon footprints, which is increasingly gaining more attention. The inclusion of household forecasts could lead to a more unified approach to the problem, the authors say. They have also highlighted that, in these modeling set-ups, limited cohesion by firm is presented as weak to accuracy by region. According to Huang and Mao[16], AI-based algorithms and data-driven methods will analyse and complement data related to global market supply chains concerning improved efficiency, cost-effectiveness, and sustainability objectives aiming at carbon footprint reduction. Basically, the novelty of this work from the perspective of AI lies in how deep learning and optimisation algorithms are applied to solve operational and environmental sustainability problems. Thus, sustainable supply chain management is promoted through this AI integration. Besides, the study further amplifies the role of AI in reducing emissions by supporting resource allocation optimisation, emissions monitoring, and real-time decision control, which benefits corporate and public policy spheres. Another research limitation appears in that it exclusively focuses on supply chain and industrial-related carbon emissions, excluding residential emissions. This becomes an essential point since the residential sector remains one of the largest contributors to global emissions, and its exclusion from the model compromises the model's completeness towards the ultimate objective of reducing emissions. This indicates a small room for gaps in literature as area edge predictions at the household level may enhance the system and offer a more thorough picture of community sustainability. Given the controversial nature of carbon emissions, this paper focuses predominantly on the study of carbon emissions as a subset only.

In another study, Liu et al.[17] prepared a high-precision, real-time, predictive model of carbon emissions from coal-fired power plants following a setting-based learning paradigm. The modeling techniques, especially the ElasticNet, were found to have a high degree of accuracy ($R^2 > 0.95$) and stability when applied to various validation scenarios. Firstly, the model analyses operational and chemical parameters in detail, identifying coal quality parameters as the major predictors,

and providing operational and strategic frameworks for emissions monitoring, optimisation, and environmental control. On the flipside, it cannot be denied that the prediction model is very much tied up to power plant data and can hardly find any use in another sector, e.g., households. The prediction of household carbon footprint remains something that has not been tackled, and this may well be an area of novelty for further research. While the industrial emission-centric study does shed some light, it also marches the landscape toward incorporating household data in future studies to push their predictive rating onto a wider, more refined level of carbon footprint management.

According to Awad and Khanna [18], the study developed an AI-based machine learning model that forecasts global CO_2 emissions for past, present, and future periods, with a view toward the consequences of the COVID-19 pandemic. It includes pandemic-related data, so it accounts for the reductions in emissions during lockdowns, and an optimised SARIMA approach is used to attain up to 91% accuracy. The data run from 2020 to 2023 and are split into pre-pandemic, start, transmission, and post-pandemic windows, allowing forecasts up to 6, 32, and 50 years ahead. Despite being very useful to policy, the limitations of the model suit include historical and pandemic-specific data only, not including external factors such as changes in policy, technology, or economic shifts that can trigger future emissions. Future works will allow improvements with the extended independent variables and optimisation methods, as well as building easy-to-use applications for real-time tracking. Overall, this novel method makes strong forecasts for emission reduction strategies while requiring further improvements for covering complicated external factors.

Garg et al. [7] performed environmental policy framing by projecting future carbon emissions in Indonesia until the year 2030 and employing machine learning-based prediction, i.e., Linear Regression. The advantage of the study rests in a straightforward forecasting model relying on historical data-well-founded enough to easily interpret and predict an upward emission trend, thus allowing policymakers to come up with informed decisions. In addition, the availability of the datasets to the public and the use of common data analysis tools add to the transparency and reproducibility of the research. On the other hand, its weaknesses involve linear regression, which might oversimplify the complex environmental and socioeconomic processes shaping emissions; it assumes a linear form along with time and, thus, might not be suited for non-linear dynamics or exogenous shocks such as social changes, technologic advances, or economic fluctuations.

Despite these advances, critical gaps remain, particularly concerning the focus of this study. A consistent limitation across the cited literature is the sectoral focus; corporate [15], supply chain [16], and industrial [18] studies explicitly exclude the residential sector, which is a major global emissions contributor. Furthermore, the major body of present literature is context-specific, aimed at developed economies with stable data infrastructures, while developing contexts such as Nigeria facing unique pressures from socioeconomic growth are largely unexplored. Finally, the predominant emphasis in existing work is on achieving predictive accuracy, with hardly any taking the element of systemic risk analysis from socioeconomic, climatic and policy-induced uncertainties into their fold. This omission limits the practical usability of such models for policy and planning where managing risk is essential.

Addressing these gaps, the present study advances literature by: (i) making use of a variety of machine learning forecasting algorithms (RFR, ETR, SVR, and XGBoost) for comparative testing; (ii) integrating predictions into a risk management framework to quantify uncertainties; and (iii) situating this analysis in a developing economy context where pressures from socioeconomic growth have to be balanced with considerations for carbon footprint reduction. It thus extends methodological rigour and offers insights that could be useful for policymakers interested in implementing SDG 13 (Climate Action) and net-zero commitments.

3. Materials and Methods

3.1. Research Methodological Framework

The research process started with a literature review to consider the key variables that influence carbon-footprint emissions across energy consumption, cli-

matic, socioeconomic, and policy-related dimensions. The most important variables were then selected and minor ones thrown out by expert judgment. In this research, inputs to the prediction models for CO_2 were considered as hours in a year, estimated income level, household electricity consumption, temperature, and household size, while outputs were considered as hourly CO_2 emission. Data collection was undertaken based on these selected variables. Hourly energy consumption data for the year 2023 were collected using a smart energy meter (**Figure 1**) installed in each of the ten (10) households (H1, H2, H3, H4, H5, H6, H7, H8, H9, and H10) at Talba estate, Minna, Niger state. Temperature data for the climatological datasets were sourced from the National Aeronautics and Space Administration (NASA) database for 2023. Other datasets such as income level and household size were collected via survey interviews with each household. Hourly CO_2 emissions for one year were measured with a developed real-time CO_2 monitoring device (**Figure 1a,b**) which was placed in each of the ten households. This yielded 87,600 hourly records (10 households × 8760 h). Fewer than 0.5% of values were missing due to sensor downtime; these were addressed using linear interpolation for continuous variables and forward-fill for static ones. The forecast window for this study is hourly, with models trained and evaluated on hourly CO_2 emission data across the entire year of 2023. Predictions were generated on the same hourly resolution to capture short-term variations in household emissions. Preprocessing steps were selected to enhance model performance: min-max normalization preserved feature distributions, z-score standardization reduced outlier influence, and k-means clustering (k = 3) identified typical consumption patterns for use as an auxiliary feature. Data was split chronologically using 70/30 ratio, with training from January to August 2023 and testing from September to December 2023 to evaluate temporal generalization. The results of the processed data were further utilized to develop the CO_2 prediction models in the subsequent section.

(**a**) (**b**)

Figure 1. Monitoring Device: (**a**) smart energy meter; (**b**) developed real-time CO_2 monitor.

3.2. Software and Implementation

All analyses were conducted in MATLAB 2023a using the statistics and machine learning toolbox. Hyperparameter tuning was performed using Bayesian optimization via the bayesopt function. Preprocessing, model training, and risk assessment were scripted in MATLAB; all code is available upon request.

3.3. Model Formulation

In this study, four machine learning models are utilized to predict carbon emissions for the considered case study; they include Random Forest Regressor, Support Vector Regressor, Extreme Gradient Boosting (XGBoost), and Extra Trees Regressor.

3.3.1. Random Forest Regressor

Random Forest is an ensemble learning method that generates a collection of decision trees during training and combines their outputs, typically by averaging, to achieve more stable predictions and limit overfitting[17,19]. Each tree is trained on a bootstrap sample of the dataset, while the choice of features at every split is randomized to promote diversity within the forest. The overall prediction function for the model is summarized

in Equation (1).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (1)$$

Where $\hat{y}$ is the final predicted output of the random forest for the input *x*.

T is the total number of decision trees.

$h_t(x)$ the prediction of the t^{th} decision tree for input *x*.

The model hyperparameters, such as the number of trees, maximum tree depth, and minimum samples per leaf, are optimized so as to achieve the best trade-off between the bias and the variance. In this study, the RFR was implemented to model household CO_2 emissions by aggregating predictions from 200 decision trees, each trained on bootstrap samples of the hourly dataset. To prevent overfitting to temporal patterns, the maximum tree depth was limited to 15, and the minimum samples per leaf was set to 5. Feature randomness at each split was controlled by sampling √P features (where *p* is the total number of features). These parameters were optimized via Bayesian optimization with 5-fold cross-validation to balance bias and variance for the high-resolution emission data.

3.3.2. Support Vector Regressor

Support Vector Regression also applies the principles of Support Vector Machines (SVM) to address regression problems by finding a function that normally deviates from the actual targets by at most epsilon, while maintaining maximum flatness [18,20]. The prediction function can be formulated as given by Equation (2).

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i*) K(x_i, x) + b \quad (2)$$

Where $f(x)$ is the predicted output of the SVM,

α_i , α_i* are the Lagrange multipliers,

$K(x_i, x)$ is the kernel function, and *b* is the bias term.

SVR's strength lies in its ability to capture nonlinear patterns using kernel methods. The main hyper parameters include the penalty parameter *C*, kernel type, and ee (epsilon) (insensitive loss margin), which are tuned through the cross-validation for the optimal predictive performance. In this study, the SVR was applied to capture potential nonlinear relationships between electricity consumption and CO_2 emissions. The Radial Basis Function (RBF) kernel was selected for its flexibility in modelling complex patterns. hyper parameters including the penalty parameter c = 10, kernel coefficient γ = 0.1, and ϵ–insensitive margin ϵ = 0.01 were tuned via Bayesian optimization with 30 evaluation trials. The SVR was particularly useful for smoothing hourly emission trends while maintaining sensitivity to consumption spikes.

3.3.3. Extreme Gradient Boosting (XGBoost)

It is possible to accelerate and scale gradient boosting by means of XGBoost. Decision trees are built in series so that each tree reduces the residual errors left by the previous trees. The learning process for the trees involves an objective function comprising a loss term plus a regularisation term to control complexity of the model [21,22], as given in Equation (3).

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i , \ _i) + \sum_{k=1}^{k} \Omega(f_k) \quad (3)$$

Where $l(y_i , \hat{y}_i)$ is the loss function (e.g., squared error), and $\Omega(f_k)$ is the regularisation term for the k^{th} tree. The final prediction is obtained by summing over all weak learners as presented in Equation (4).

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), \quad f_k \in \zeta \quad (4)$$

Here, f stands for the space of regression trees, and one fine-tunes all XGBoost parameters like learning rate (ζ), maximum depth of trees, subsampling ratio, and number of estimators, so as to ideally balance bias, variance, and computational complexity.

XGBoost was employed to sequentially correct residual errors in hourly emission predictions. The model was configured with 500 boosting rounds, a learning rate of 0.05, and maximum tree depth of 6 to prevent overfitting to noisy hourly data. L2 regularization (λ = 1) and subsampling (70% of data per round) were applied to enhance generalization. These parameters were optimized using Bayesian optimization with 3 fold cross-validation, specifically targeting the reduction of RMSE across all households.

3.3.4. Extra Trees Regressor

Extra Trees Regressor is a more randomized version of the Random Forest as it adds randomness on top

of what is already present during the tree-building process. Whereas in the case of Random Forest the best split is picked from a subset of features, in the case of Extra Trees instead it picks randomly the thresholds for each feature and then calculates which one is the best, thus, variance is reduced even more[23–25]. The prediction function is given by Equation (5).

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} {h_t}^{extra}(x) \quad (5)$$

Where ${h_t}^{extra}$ represents the prediction from each extremely randomised tree. An increase in diversity among trees is usually accompanied by better generalisation, at the cost of a small increase in bias. Performance is optimised by tuning hyperparameters such as the measurement being estimated, the depth of the trees, and the minimum number of samples on a split.

The Extra Trees Regressor in this study was implemented to increase model diversity through additional randomization during tree construction. Unlike Random Forest, split thresholds were chosen completely at random for each feature, reducing variance in household emission predictions. The ensemble consisted of 300 trees with no maximum depth limit, and splits were accepted only if they partitioned at least 2 samples. This configuration was particularly effective in smoothing volatile emission patterns while maintaining computational efficiency.

3.4. Prediction-Error Risk Assessment Model

The presented model of prediction-error risk assessment, illustrated in **Figure 2**, incorporates detailed analysis of model prediction errors, quantifying the uncertainty associated with carbon footprint forecasts for each household.

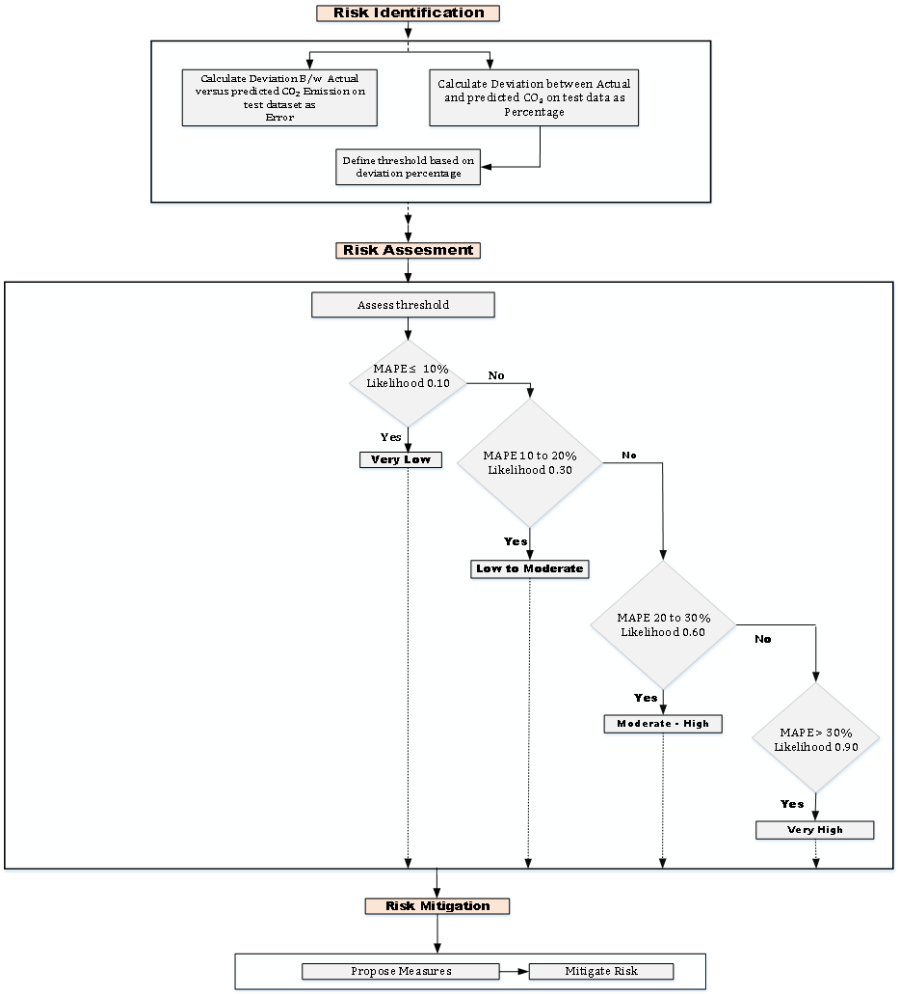


Figure 2. Prediction Error Risk Management Model of CO_2 Emission Prediction.

The likelihood thresholds (MAPE ≤ 10% = 0.10 for very low likelihood, 10% < MAPE ≤ 20% = 0.30 for low-moderate likelihood, 20% < MAPE ≤ 30% = 0.60 for moderate-high likelihood, and MAPE > 30% = 0.90 for very high likelihood), aligns with forecasting performance benchmarks [26–28]. The impact dimension was calculated as MAE divided by mean observed emissions, capped at 1. The final risk categories (Low, Moderate, High, Severe) follows ISO 31000 risk matrix conventions [29,30].

The model begins by computing standard performance metrics, including Mean Absolute Percentage Error (MAPE), Mean Absolute Error [31], Root Mean Square Error (RMSE) and the coefficient of determination (R^2) statistics were calculated for every single household. MAPE's values were then placed on a scale of probabilities, where MAPE ≤ 10% was assigned a likelihood of 0.10 (very low), 10–20% to 0.30 (low-moderate), 20–30% to 0.60 (moderate-high) and anything above 30% was marked as 0.90 (very high). Prediction error was measured in terms of its effect on MAE being proportionate to the average observed CO_2 emissions, with a limit of 1 set to prevent overstating the impact of prediction errors. The product of likelihood and impact is then scaled to generate a risk score (0–100), which is categorised into four risk levels: Low (≤10), Moderate (11–30), High (31–60), and Severe (>60). Each risk level is assigned tailored mitigation strategies, ranging from periodic monitoring and frequent validation with hyperparameter tuning (Moderate) to retraining with more data and feature engineering (High), and immediate review with fallback rules (Severe).

3.5. Model Parameter Selections

Random Forest regression was implemented using Bagging ('Method', 'Bag') with decision trees as base learners, while Extra Trees regressor also used Bagging but introduced additional randomness by sampling feature subsets to enhance generalisation; in both cases, Bayesian optimisation tuned NumLearningcycles (trees). To balance exploration and exploitation, the function expected-improvement-plus was used together with MinLeafSize (leaf size/complexity), and MaxNumSplits or NumVariablesToSample (feature-level randomness/depth). In the case of SVM, Bayesian optimisation was preferred over grid or random search because of its effectiveness in searching through intricate hyperparameter settings. With 30 evaluations, 5-fold cross-validation, and suppressed plotting/output to ensure stability, computational efficiency, and robust evaluation under an 80/20 train-test split, this approach was employed. In XGBoost, NumLearningCycles (boosting rounds), LearnRate (step size), MinLeafSize, and MaxNumSplits (tree depth control) were tuned through Bayesian optimisation using 30 trials and 3-fold cross-validation employing the same acquisition function, ensuring minimized prediction error, improved stability, and strong generalisation across households.

3.6. Model Evaluation

The fouling rate curves predicted are compared with the measured ones to evaluate the accuracy of the model. Besides graphical representations, numerical statistical indices are used to present the predictions. Among the statistical estimates are the Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Absolute Error [31], and the Coefficient of Determination (R^2) [1,31], described in Equations (6)–(9). The lower the values of MAPE, RMSE, and MAE, the more successful the prediction is. On the other hand, an R^2 value that tends to 1 indicates a very fine correlation among the predicted and observed values. Taken together, these measures offer a balanced and reliable evaluation of the model's predictive strength.

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{6}$$

$$\text{MSE} = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{9}$$

Where n = number of test samples.

y_i = actual (observed) CO_2 emission value for the i^{th} sample.

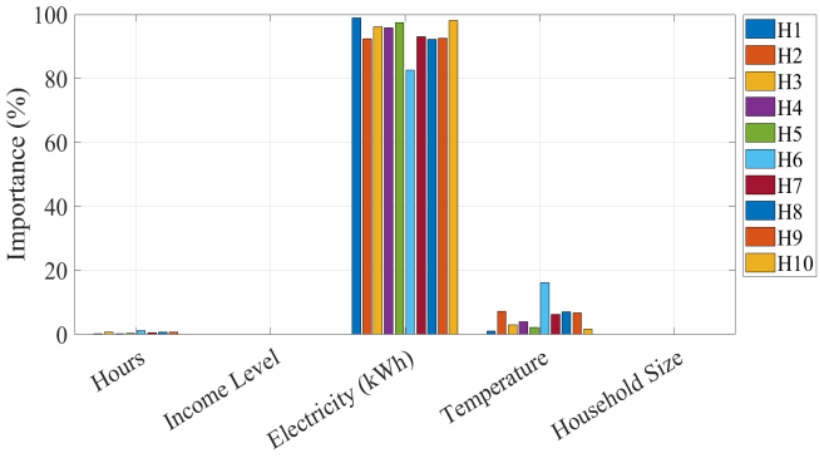
$\hat{y}_i$ = predicted CO_2 emission value for the i^{th} sample.

4. Results

4.1. Feature Importance

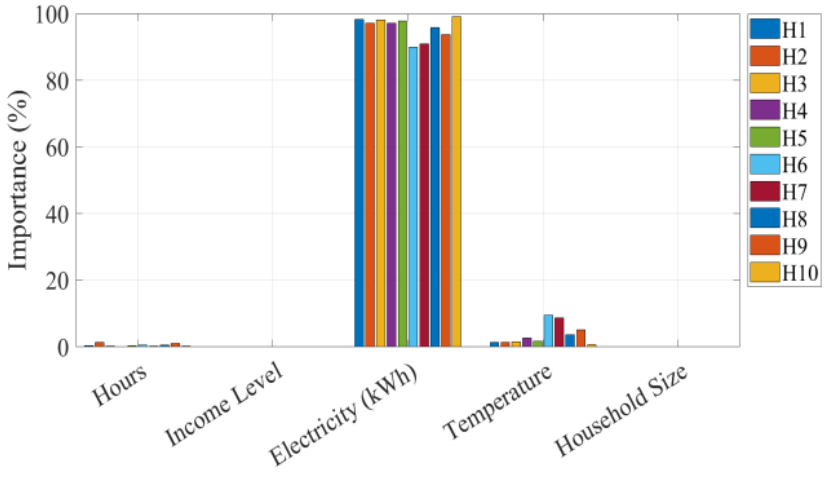
The feature importance of the four models is shown in **Figure 3**. Upon analyzing the Random Forest Regressor, it was found that electricity consumption (in kWh) is the main determinant of CO_2 emissions for all households, depending on the scenario, between 80% and 95% of the total predictor's importance. For some households, temperature may be regarded as a secondary factor that indicates climate-driven energy requirements for heating or cooling. Household size, income level, and hours have a very slight effect, which means, from a predictive point of view, these factors have a minor impact on electricity usage. Overall, reducing electric power usage is the best way to reduce household CO_2 emissions. The Extra Trees Regressor model asserts that kWh electricity consumption is the one and only predictor of CO_2 emission for the ten households, taking almost all the importance (100%). Other characteristics, including hours, income level, temperature, and household size, have insignificant effects, implying that they cannot add any predictive value when considering electricity use. The analysis based on the Support Vector Regressor (SVR) demonstrates that the consumption of electricity (kWh) is the most important predictor of the amount of CO_2 emitted by all houses, accounting for more than 80–95% of the weight. In certain households, temperature plays a minor role, corresponding to the energy requirements of the Climate. The household size, income level, and hours do not significantly impact or improve the predictive ability, in addition to the usage of electricity. Altogether, SVR focuses on electricity as the most influential source of emissions, with a negligible consideration of temperature impacts. The XGBoost Regressor analysis indicates that electricity consumption (kWh) is the only predictor that dominates CO_2 emissions in all ten households, accounting for almost 100 per cent of the importance. The other characteristics, such as hours, income level, temperature, and household size, have only a minimal impact, meaning they do not significantly contribute to the predictive value of electricity use when considered.

The findings are in agreement with those of all households, indicating the high dependence of the model on direct electricity data. This implies that we can precisely predict the CO_2 emissions of households by observing their electricity consumption alone, but secondary factors are often overlooked. This is physically grounded, as the primary source of emissions for these households is grid electricity, whose carbon intensity is directly proportional to consumption. The negligible importance of temperature, income, and household size is likely due to the study's context; in the tropical climate of Minna, temperature-driven heating/cooling loads are minimal, and income variations within the sampled estate may not be large enough to cause significant differences in appliance ownership or usage intensity that aren't already captured by the kWh reading. This finding simplifies the monitoring premise, suggesting that in similar off gas-grid, tropical urban settings, smart meter data alone may be a robust proxy for household carbon footprint.

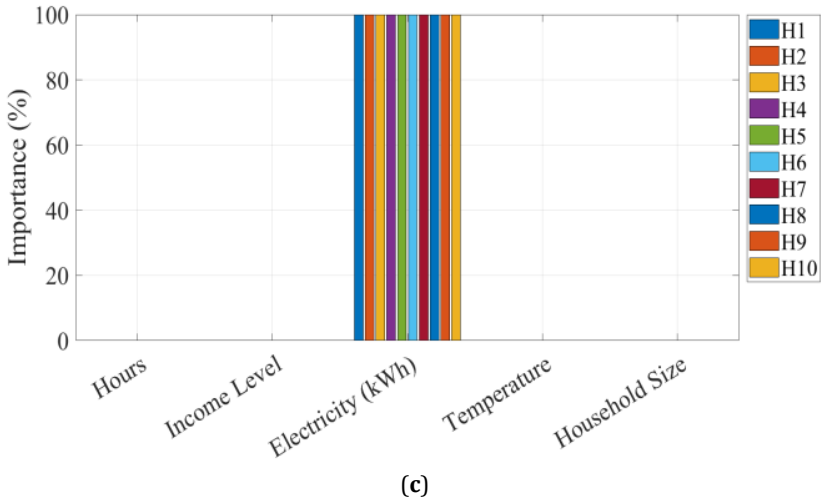


(**a**)

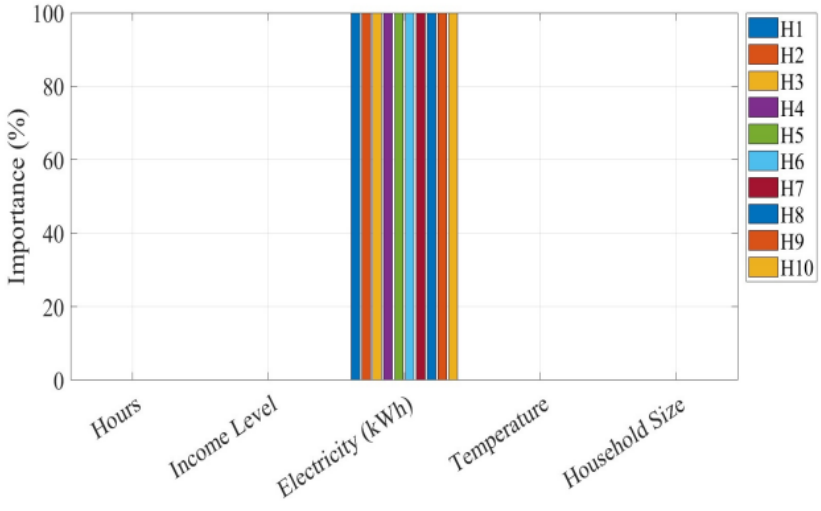
Figure 3. *Cont.*



(**b**)



(**c**)

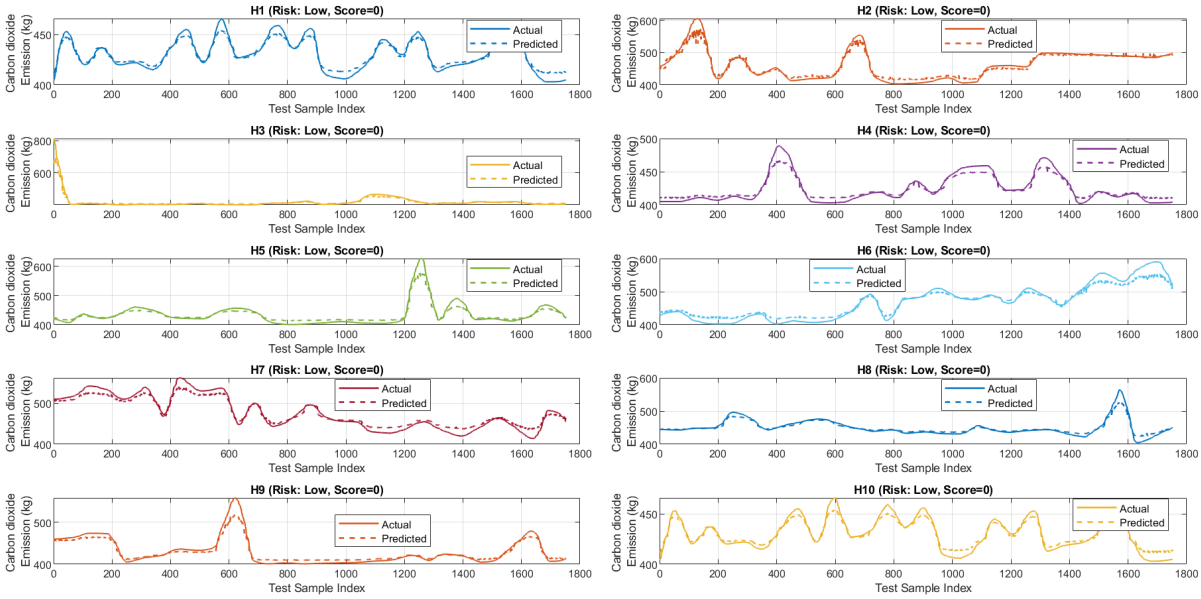


(**d**)

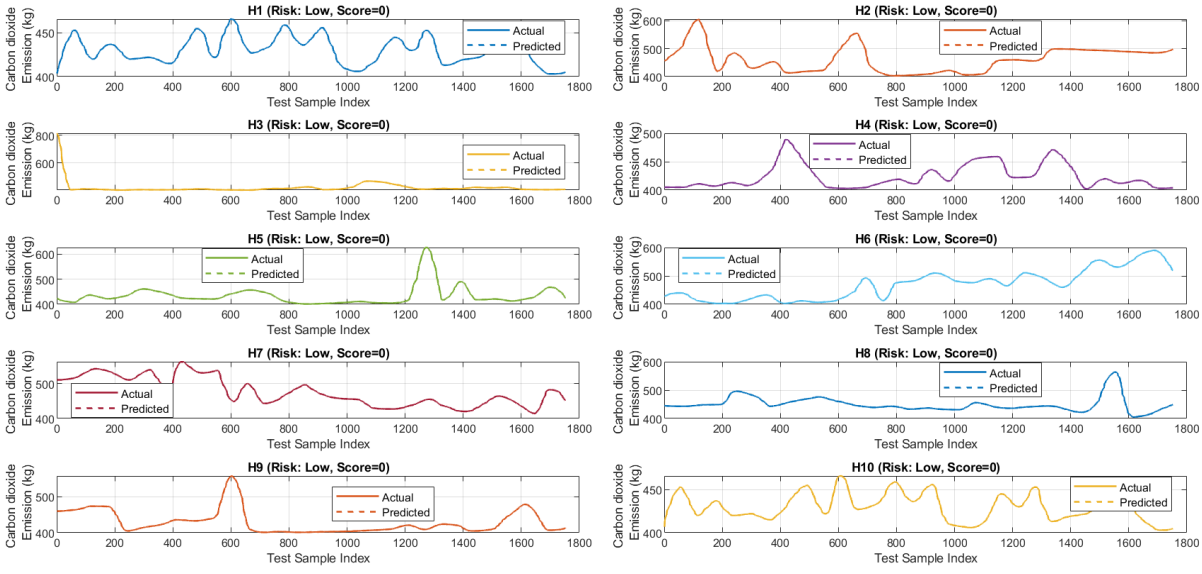
Figure 3. Feature Importance for the Four (4) Prediction Models: (**a**) Random Forest; (**b**) Extra Trees Regressors; (**c**) Support Vector Regressors; (**d**) XGBoost.

4.2. Actual versus Predicted CO_2 Emission on Test Datasets

As shown in **Figure 4a–d**, for Talba Estate, there is a variance with their accuracies with respect to CO_2 emission predictions (based upon-date entered): Random Forest Regressor, Extra Tree Regressor, Support Vector Regressor (SVR), and XGBoost. Models perform well upon smooth or moderate emissions (e.g., H1, H3, H5, H6, H8) but are unable to manage the sharp peaks (e.g., H2, H4, H7). SVR underestimates sudden spikes, Extra Trees smooth variations, Random Forest effectively captures trends with some deviations, and XGBoost closely tracks peaks with minor underestimations. Households 9 and 10 show consistent tracking with slight smoothing. Together, XGBoost and Random Forest are most suitable for dynamic fluctuations, Extra Trees for trend estimations, and SVR in stable conditions but less effectively in volatile ones.

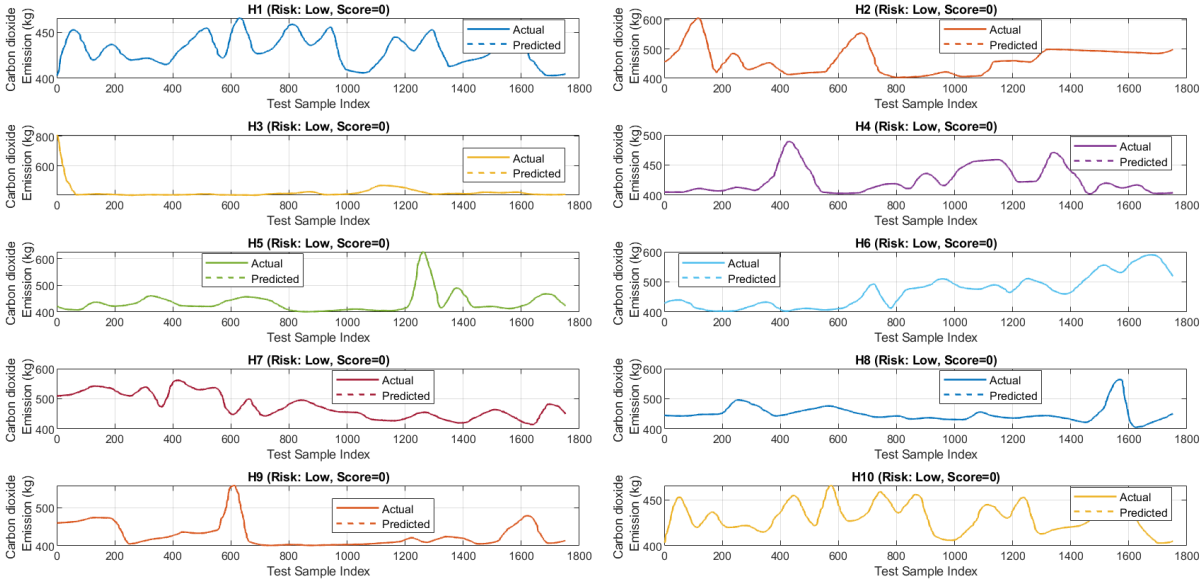


(a)

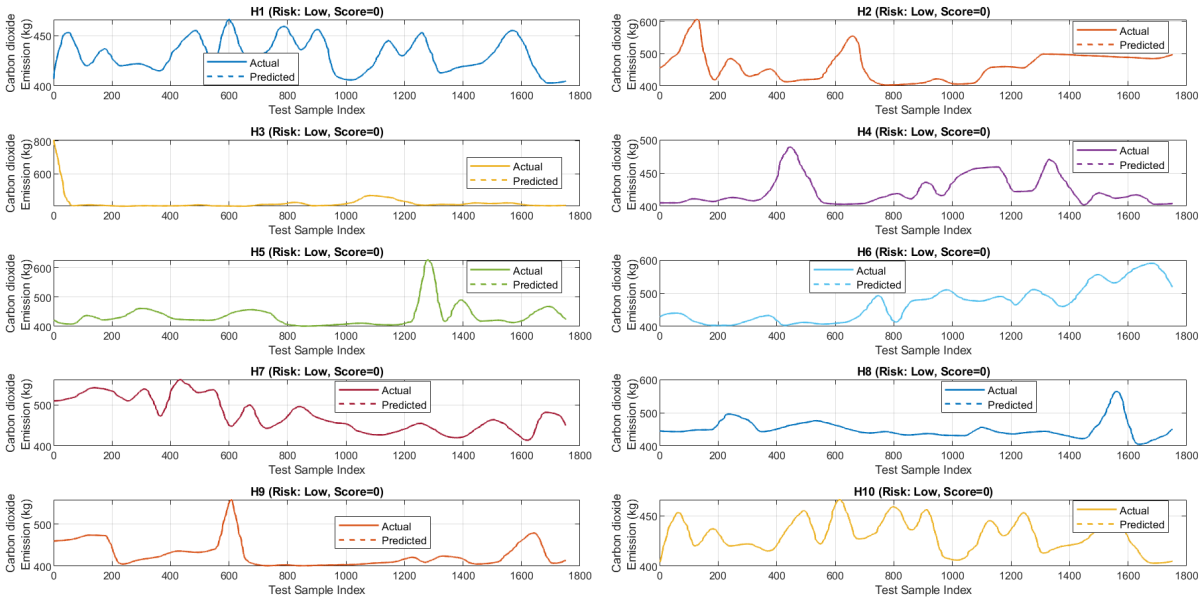


(b)

Figure 4. *Cont.*



(c)



(d)

Figure 4. Actual Versus Predicted Carbon dioxide Emission for the Four (4) Prediction Models: (**a**) Random Forest; (**b**) Extra Trees Regressors; (**c**) Support Vector Regressors; (**d**) XGBoost.

The superior performance of XGBoost and tree based ensembles (ETR and RFR) in capturing emission spikes (e.g., in H2, H4 and H7) can be attributed to their inherent ability to model non-linear, threshold-based behaviors; such as the simultaneous activation of high-wattage appliances. Conversely, SVR's tendency to smoot sharp peaks aligns with its formulation around an epsilon-insensitive loss function, which treats minor deviations within a margin as zero error. The higher error metrics for some households (e.g., H5, H6) likely reflects more irregular, occupant-dependency activity patterns that are less predictable from consistent variables like hourly timestamp, highlighting the limit of models based solely on scheduled or meteorological factors.

4.3. Model Evaluation Results

The comparative performance analysis across the four machine learning models: Random Forest Regressors (RFR), Extra Trees Regressors (ETR), Support Vec-

tor Regressors (SVR), and XGBoost, is presented in **Table 1**.

The ranking of performance is very evident in the results, with ETR, SVR, and XGBoost being the top three. The Random Forest Regressor, however, is RFR that still managed to achieve moderate prediction accuracy with overall MAPE of 1.34%, RMSE of 9.05, and R^2 of 0.9472. This shows a good fitting but still makes improvement possible, especially in the cases of households with higher load variability such as H6 and H5, which have comparatively higher RMSE and MSE values. ETR, on the other hand, gives a major boost in prediction performance and thus the overall MAPE is drastically reduced to 0.0160, RMSE to 0.1583, and R^2 is almost perfect of 0.9999, showing remarkable generalization indeed. In addition, SVR provides support with an overall MAPE of only 0.0243 and a highly proximate R^2 of 0.9999, especially in the aforementioned households H1, H3, H4, H6, H8, and H9, where the errors in the predictions are negligible. Yet, the case of H7 is different as it shows a fairly high SVR error (MAPE = 0.1231, RMSE = 0.7164), which signals that it is easily affected by data irregularities. XGBoost, on the other hand, is the best ranked overall by rightly combining very low error metrics (overall MAPE = 0.0073, RMSE = 0.1463, MAE = 0.0340) along with perfect predictive power (R^2 ≈ 1.0000) and is thus the model that has consistently outrun the others in all the households. Overall, while RFR provides a decent baseline, ETR, SVR, and especially XGBoost demonstrate superior predictive accuracy and robustness, making XGBoost the most reliable model for predicting CO_2 emissions in this study.

Table 1. Results of the models' evaluation using MAPE, RMSE, MAE and R^2.

Households (H)	MAPE	RMSE	MAE	MSE	R^2	MAPE	RMSE	MAE	MSE	R^2
Random Forest Regressors						**Extra Trees Regressors**				
H1	0.7716	4.1739	3.3165	17.4210	0.9278	0.0061	0.0375	0.0262	0.0014	0.9999
H2	1.5889	10.3700	7.3802	107.5300	0.9467	0.0195	0.1194	0.0883	0.0142	0.9999
H3	0.9791	9.9269	4.5129	98.5430	0.9353	0.0235	0.3895	0.1097	0.1517	0.9999
H4	1.1888	6.5676	5.1353	43.1340	0.9062	0.0138	0.0784	0.0586	0.0061	0.9999
H5	1.7262	10.7570	7.7719	115.7100	0.9165	0.0211	0.1499	0.0930	0.0225	0.9999
H6	2.2211	13.7600	10.5490	189.3400	0.9298	0.0248	0.1524	0.1160	0.0232	0.9999
H7	1.7675	10.1420	8.4045	102.8700	0.9372	0.0177	0.1151	0.0847	0.0132	0.9999
H8	0.9664	6.9219	4.4275	47.9120	0.9160	0.0109	0.0774	0.0499	0.0060	0.9999
H9	1.4145	8.8158	6.2937	77.7180	0.9251	0.0171	0.1030	0.0746	0.0106	0.9999
H10	0.8044	4.3729	3.4516	19.1220	0.9216	0.0061	0.0384	0.0260	0.0015	0.9999
Overall	1.3429	9.0515	6.1243	81.9300	0.9472	0.0160	0.1583	0.0727	0.0251	0.9999
Support Vector Regressors						**XGBoost**				
H1	0.0041	0.0366	0.0177	0.0013	0.9999	0.0025	0.0175	0.0110	0.0003	1.0000
H2	0.0520	0.2841	0.2393	0.0807	0.9999	0.0049	0.0433	0.0238	0.0019	1.0000
H3	0.0046	0.0288	0.0199	0.0008	1.0000	0.0233	0.4317	0.1142	0.1864	0.9999
H4	0.0022	0.0106	0.0093	0.0001	1.0000	0.0175	0.1147	0.0754	0.0131	0.9999
H5	0.0218	0.1103	0.0941	0.0122	0.9999	0.0048	0.0662	0.0231	0.0044	1.0000
H6	0.0110	0.0744	0.0525	0.0055	1.0000	0.0033	0.0337	0.0163	0.0011	1.0000
H7	0.1231	0.7164	0.5842	0.5132	0.9997	0.0044	0.0316	0.0211	0.0010	1.0000
H8	0.0159	0.0760	0.0705	0.0058	0.9999	0.0036	0.0430	0.0171	0.0018	1.0000
H9	0.0027	0.0133	0.0115	0.0002	1.0000	0.0057	0.0603	0.0261	0.0036	1.0000
H10	0.0051	0.0261	0.0221	0.0007	1.0000	0.0027	0.0164	0.0117	0.0003	1.0000
Overall	0.0243	0.2491	0.1121	0.0620	0.9999	0.0073	0.1463	0.0340	0.0214	0.9999

4.4. Risk Management Results

The risk management results for all four models: Random Forest Regressor, Extra Trees Regressor, Support Vector Regressor, and XGBoost, are presented in **Table 2**.

The overall risk assessments reveal steady low-risk positions all the time for every household (H1–H10) which is the way the small risk scores (all practically 0.0000 after rounding) indicate them. Regarding the likelihood values, they are still 0.100 for every household, and then the impact values are so small that it is hard to see them except for slight changes among the households in the Random Forest and Extra Trees models but not at all in the SVR and XGBoost models where it is almost impossible to notice them. One can, therefore, conclude that there is an extremely low chance of any significant prediction error across all the used ma-

chine learning methods. The chosen mitigation strategy (MMPP: Monitor Models Performance Periodically) is enforced uniformly across the board, thereby guaranteeing ongoing evaluation of the models for accuracy maintenance over a long period. In general, the results suggest that the four models are very reliable, resilient, and not a potential source of error in the CO_2 estimation and projection process of any household.

Table 2. Risk management result.

H	Likelihood	Impact	RiskScore	Risk Level	Mitigation	Likelihood	Impact	RiskScore	Risk Level	Mitigation
Random Forest Regressors						**Extra Trees Regressors**				
H1	0.100	7.7×10^{-3}	7.7×10^{-4}	Low	MMPP	0.100	2×10^{-5}	2×10^{-6}	Low	MMPP
H2	0.100	1.6×10^{-2}	1.6×10^{-3}	Low	MMPP	0.100	1.8×10^{-4}	1.8×10^{-5}	Low	MMPP
H3	0.100	1.07×10^{-2}	1.07×10^{-3}	Low	MMPP	0.100	2.8×10^{-4}	2.8×10^{-5}	Low	MMPP
H4	0.100	1.21×10^{-2}	1.21×10^{-3}	Low	MMPP	0.100	1.2×10^{-4}	1.2×10^{-5}	Low	MMPP
H5	0.100	1.79×10^{-2}	1.79×10^{-3}	Low	MMPP	0.100	1.7×10^{-4}	1.7×10^{-5}	Low	MMPP
H6	0.100	2.25×10^{-2}	2.25×10^{-3}	Low	MMPP	0.100	1.6×10^{-4}	1.6×10^{-5}	Low	MMPP
H7	0.100	1.76×10^{-2}	1.76×10^{-3}	Low	MMPP	0.100	1.9×10^{-4}	1.9×10^{-5}	Low	MMPP
H8	0.100	9.8×10^{-3}	9.80×10^{-4}	Low	MMPP	0.100	1.1×10^{-4}	1.1×10^{-5}	Low	MMPP
H9	0.100	1.47×10^{-2}	1.47×10^{-3}	Low	MMPP	0.100	1.95×10^{-4}	1.95×10^{-5}	Low	MMPP
H10	0.100	8.0×10^{-3}	8.0×10^{-4}	Low	MMPP	0.100	1.5×10^{-5}	1.5×10^{-6}	Low	MMPP
Support Vector Regressors						**XGBoost**				
H1	0.1000	4.2×10^{-5}	4.2×10^{-6}	Low	MMPP	0.1000	3.1×10^{-5}	3.1×10^{-6}	Low	MMPP
H2	0.1000	5.1×10^{-4}	5.1×10^{-5}	Low	MMPP	0.1000	4.5×10^{-5}	4.5×10^{-6}	Low	MMPP
H3	0.1000	3.8×10^{-5}	3.8×10^{-6}	Low	MMPP	0.1000	2.75×10^{-4}	2.75×10^{-5}	Low	MMPP
H4	0.1000	2.9×10^{-5}	2.9×10^{-6}	Low	MMPP	0.1000	1.8×10^{-4}	1.80×10^{-5}	Low	MMPP
H5	0.1000	1.95×10^{-4}	1.95×10^{-5}	Low	MMPP	0.1000	2.5×10^{-5}	2.5×10^{-6}	Low	MMPP
H6	0.1000	1.05×10^{-4}	1.05×10^{-5}	Low	MMPP	0.1000	3.5×10^{-5}	3.5×10^{-6}	Low	MMPP
H7	0.1000	1.2×10^{-3}	1.2×10^{-4}	Low	MMPP	0.1000	4×10^{-5}	4×10^{-6}	Low	MMPP
H8	0.1000	1.75×10^{-4}	1.75×10^{-5}	Low	MMPP	0.1000	2.2×10^{-5}	2.2×10^{-6}	Low	MMPP
H9	0.1000	3.5×10^{-5}	3.50×10^{-6}	Low	MMPP	0.1000	3×10^{-5}	3×10^{-6}	Low	MMPP
H10	0.1000	2.5×10^{-5}	2.5×10^{-6}	Low	MMPP	0.1000	2.8×10^{-5}	2.8×10^{-6}	Low	MMPP

Note: MMPP = Monitor Model Performance Periodically, H= Household.

5. Conclusions

This investigation reveals that the machine learning models developed, particularly XGBoost, can accurately predict household carbon dioxide emissions in developing country contexts such as Nigeria, even with data obtained from only ten households. Very low error values were obtained from the models which indicate that their predictions are very stable and very uncertain in spite of the dynamic changes. The consistently "Low Risk" scores across all households and models are a direct mathematical consequence of the exceptionally low prediction errors (MAPE, MAE) achieved, particularly by the top-performing models. This does not indicate a lack of sensitivity in the risk metric but rather validates the core robustness and reliability of the forecasting framework for this dataset. The near-zero impact values confirm that prediction errors are minimal relative to the actual emission scales. For policymakers, this low-risk outcome is meaningful; it builds confidence that data-driven models can provide stable baselines for monitoring and targeting emission reductions at the household level. A critical future step is to stress-test this framework with noisier data or in less predictable settings to define the boundaries of its low-risk applicability.

The above-mentioned results point out the great potential inherent in pairing advanced predictive algorithms with prediction-error risk management frameworks to improve confidence in household CO_2 forecasts. The integration of high-accuracy prediction with a formal risk assessment provides a dual tool for climate action. Technically, it demonstrates that machine learning can demystify household emissions, turning them from an estimated average into a predictable, managed variable. For policy, this enables targeted strategies: utilities or regulators could use such a framework to identify high-emitting households for tailored efficiency programs or to reliably certify emission reductions for carbon credit initiatives. The minimal role of socioeconomic variables in our model suggests that, in the

near term, direct interventions on electricity consumption through efficient appliances and renewable energy integration would yield the most measurable impact, directly supporting SDG 13 targets and net-zero community planning.

The findings, while accurate, are constrained by the small and geographically narrow dataset (10 households in one estate) and the exclusion of external variables like policy shifts or technological adoption rates. The findings propose that policymakers and utilities could integrate such a predictive-risk framework into national carbon monitoring systems and design targeted incentive or retrofit programs based on household emission profiles identified by the model, directly supporting SDG 13 and net-zero planning. The study addresses the limitations for future work by suggesting expanding the data cohort across diverse Nigeria climatic and socioeconomic zones, incorporating behavioral survey data on appliance use, and developing user-friendly dashboard platforms to enable real-time use by households and energy planners.

Author Contributions

Idea and conceptualization, research and investigation, analysis, methodology, review and editing, supervision, funding acquisition, project administration, L.O.; research and investigation, analysis, methodology, data curation, visualization, software and simulation, original draft preparation, review and editing, H.O.I.; research analysis and investigation, resources, writing—review and editing, supervision, project administration, J.G.A.; research analysis and investigation, resources, writing—review and editing, supervision, project administration, U.S.D.; writing—review and editing, I.S.; writing—review and editing, supervision, K.E.J. All authors have read and agreed to the published version of the manuscript.

Funding

This study received support from the Tertiary Education Trust Fund (TETFund), Nigeria, through the National Research Fund (NRF) intervention, under project ID: TETF/ES/DR&D/CE/NRF2021/CC/CAE/00114.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data will be made available on request.

Acknowledgments

Our gratitude goes to household members of Talba Estate for their support in conducting this research and also to Tertiary Education Trust Fund (TETFund), Nigeria for their support and financial assistance in providing funding.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Dong, Q., Zhong, C., Geng, Y., et al., 2024. A bibliometric review of carbon footprint research. Carbon Footprints. 3, 3. DOI: https://dx.doi.org/10.20517/cf.2023.45

[2] Yuan, R., Rodrigues, J.F., Wang, J., et al., 2022. A global overview of developments of urban and rural household GHG footprints from 2005 to 2015. Science of the Total Environment. 806(Part 2), 150695. DOI: https://doi.org/10.1016/j.scitotenv.2021.150695

[3] Nwokolo, S.C., Meyer, E.L., Ahia, C.C., 2023. Credible pathways to catching up with climate goals in Nigeria. Climate. 11(9), 196.

[4] Mudhee, K.H., Hilal, M.M., Alyami, M., et al., 2025. Assessing climate strategies of major energy corporations and examining projections in relation to Paris Agreement objectives within the framework of sustainable energy. Unconventional Resources. 5, 100127.

[5] Teske, S., 2019. Achieving the Paris Climate Agreement Goals: Global and Regional 100% Renewable Energy Scenarios with Non-Energy GHG Pathways for +1.5 °C and +2 °C. Springer Nature:

Cham, Switzerland.

[6] Meinshausen, M., Lewis, J., McGlade, C., et al., 2022. Realization of Paris Agreement pledges may limit warming just below 2 °C. Nature. 604(7905), 304–309. DOI: https://doi.org/10.1038/s41586-022-04553-z

[7] Garg, A.P., Chaudhary, M., Garg, C., 2024. Global impact of carbon emissions and strategies for its management. In Quality of Life and Climate Change: Impacts, Sustainable Adaptation, and Social-Ecological Resilience. IGI Global Scientific Publishing: Hershey, PA, USA. pp. 75–107.

[8] Amin, R., Ar Salan, M.S., Hossain, M.M., 2024. Measuring the impact of responsible factors on CO_2 emission using generalized additive model (GAM). Heliyon. 10(4), e25416. DOI: https://doi.org/10.1016/j.heliyon.2024.e25416

[9] Alli, Y.A., Bamisaye, A., Bamidele, M.O., et al., 2024. Transforming waste to wealth: Harnessing carbon dioxide for sustainable solutions. Results in Surfaces and Interfaces. 17, 1–34.

[10] Hoa, P.X., Xuan, V.N., Thu, N.T.P., 2024. Factors affecting carbon dioxide emissions for sustainable development goals—New insights into six Asian developed countries. Heliyon. 10(21), e39943.

[11] Jin, Y., Sharifi, A., Li, Z., et al., 2024. Carbon emission prediction models: A review. Science of the Total Environment. 927, 172319.

[12] Li, Q., Jia, R., Du, Q., et al., 2025. Drivers and multi-scenario projections of life cycle carbon emissions from China's construction industry. Sustainability. 17(9), 3828. DOI: https://doi.org/10.3390/su17093828

[13] Müller, L.J., Kätelhön, A., Bachmann, M., et al., 2020. A guideline for life cycle assessment of carbon capture and utilization. Frontiers in Energy Research. 8, 1–20.

[14] Nguyen, V.G., Duong, X.Q., Nguyen, L.H., et al., 2023. An extensive investigation on leveraging machine learning techniques for high-precision predictive modeling of CO_2 emission. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects. 45(3), 9149–9177.

[15] Nguyen, Q., Diaz-Rainey, I., Kuruppuarachchi, D., 2021. Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach. Energy Economics. 95, 1–18.

[16] Huang, R., Mao, S., 2024. Carbon footprint management in global supply chains: A data-driven approach utilizing artificial intelligence algorithms. IEEE Access. 12, 89957–89967.

[17] Liu, L., Liu, J., Zhu, X., 2024. An augmentation of random forest through using the principle of justifiable granularity. In Proceedings of the Third International Conference on Electronic Information Engineering, Big Data, and Computer Technology (EIBDCT 2024), Beijing, China, 26–28 January 2024; pp. 1316–1324.

[18] Awad, M., Khanna, R., 2015. Support vector regression. In Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Springer: London, UK. pp. 67–80.

[19] Pious, I.K., Rajalakshmi, A., Kumar, P., et al., 2024. Enhancing prediction accuracy through random forest in classification and regression. In Proceedings of the 2024 International Conference on Smart Technologies for Sustainable Development Goals (ICSTSDG), Chennai, India, 6–8 November 2024; pp. 1–6.

[20] Orchel, M., 2011. Support vector regression as a classification problem with a priori knowledge in the form of detractors. In Man–Machine Interactions 2. Springer: Berlin, Germany. pp. 351–358.

[21] Fafalios, S., Charonyktakis, P., Tsamardinos, I., 2020. Gradient boosting trees. Gnosis Data Analysis PC. 1, 1–3.

[22] Mitchell, R., Adinets, A., Rao, T., et al., 2018. XGBoost: Scalable GPU accelerated learning. arXiv preprint. arXiv:1806.11248. DOI: https://doi.org/10.48550/arXiv.1806.11248

[23] Alshare, S., Abdullah, M., Quwaider, M., 2022. Increasing accuracy of random forest algorithm by decreasing variance. In Proceedings of the 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 21–23 June 2022; pp. 232–238.

[24] Sharma, R.K.M., Unnisa, S., 2024. Airline price prediction using ExtraTrees regressor. International Journal of Information Technology, Research and Applications. 3(3), 7–14.

[25] Mastelini, S.M., Nakano, F.K., Vens, C., et al., 2022. Online extra trees regressor. IEEE Transactions on Neural Networks and Learning Systems. 34(10), 6755–6767.

[26] Kim, S., Kim, H., 2016. A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting. 32(3), 669–679.

[27] Tofallis, C., 2015. A better measure of relative prediction accuracy for model selection and model estimation. Journal of the Operational Research Society. 66(8), 1352–1362.

[28] de Myttenaere, A., Golden, B., Le Grand, B., et al., 2016. Mean absolute percentage error for regression models. Neurocomputing. 192, 38–48.

[29] Nordal, Y.A.B., 2020. A simple scenario-based qualitative model for assessing start-up risks. In Proceedings of the 2nd International Conference

on Finance, Economics, Management and IT Business (FEMIB 2020), online, 5–6 May 2020; pp. 98–105.

[30] Jenke, D., 2015. Safety risk categorization of organic extractables associated with polymers used in packaging, delivery and manufacturing systems for parenteral drug products. Pharmaceutical Research. 32(3), 1105–1127.

[31] Afrifa, G.A., Tingbani, I., Yamoah, F., et al., 2020. Innovation input, governance and climate change: Evidence from emerging countries. Technological Forecasting and Social Change. 161, 1–41.