International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 14, Issue 1, 2026

Carbon Footprint Estimation Using Ensemble Learning Techniques

R Dinesh Kumar[1], N Satya Srinija [2], A S C L S Sruthi[3], G Sohana[4]

[1]Assistant Professor;Department of CSE, Bhoj Reddy Engineering College For Women, Hyderabad, India

[2,3,4]B.Tech Students; Department of CSE, Bhoj Reddy Engineering College For Women, Hyderabad, India

Mailid;me.dineshkumar@gmail.com[1],narayanamsrinija@gmail.com[2],asruthi3525@gmail.com[3],gsohana95@gmail.com[4]

Abstract

Carbon footprint must be accurately estimated to ensure sustainable practices are encouraged as well as assist environmental decisions making. In recent years, the methods of machine learning have proven useful in predicting and analyzing emissions, and this is due to the growing access to large-scale environmental and transactional data. The paper is an estimation framework of carbon footprint using ensemble learning methods to enhance the accuracy and robustness of the prediction. The approach under proposal incorporates the use of several base learners such as decision trees, random forests, and gradient boosting models to learn complicated nonlinear relationships between energy consumption, transportation patterns, and production activities. To improve the model performance and minimize noise, the feature selection and data preprocessing methods are used. The ensemble model is trained and tested on real world data and the performance of the ensemble model against that of the individual machine learning models is compared through the use of standard evaluation measures like the mean absolute error and the root mean square error. It has been proven by experiments that the ensemble-based approach is more accurate and stable in estimating carbon emissions. The suggested system is applicable successfully to smart cities, industrial surveillance, and e-commerce infrastructures to facilitate a sustainable development and carbon-reduction policy.

Keywords: *Carbon footprint, Ensemble learning, Machine learning, Sustainability, Deep learning.*

Introduction

Climate change has become one of the major concerns facing the world as a whole because of the relentless increase in the levels of greenhouse gases (GHG) especially carbon dioxide (CO2) through industrial, transportation, agricultural and domestic practices. Proper determination of carbon footprint, which is the sum total of GHGs emissions that can in one way or the other be linked to an activity, a product or an organization or the individual, is vital in mitigation strategies and in attaining a sustainable agenda.

Standard carbon footprint assessment processes are based on manual computation, emission factor and fixed statistical models. Even though such methods give some basic estimates, they cannot usually account for more intricate interactions between various influencing parameters and dynamic consumption patterns. Furthermore, the traditional models can be limited in their scalability in cases when used with large and non-homogeneous data. As more big data is available in the form of smart devices, industrial systems and digital platforms, there has been a rise in data-driven methods as possible alternatives in the analysis of carbon emissions.

Machine learning methods allow learning automatically through past data and the identification of the latent patterns affecting the emission of carbon. Problem solving models (PS) Single problem solving models, e.g., decision trees, support vector machines or linear regression, can be limited in their generalization performance when presented with high-dimensional or noisy data. The models are also sensitive to imbalance and outliers of the data which may influence the prediction reliability. These limitations have been addressed by the attention on ensemble learning to enhance the accuracy, stability, and robustness of models.

Ensemble learning uses several base models to achieve a single prediction result thus minimizing the individual model bias and variance. Ensemble approaches are in a better position to manage complex nonlinear relationships in the environmental and consumption data by combining the strengths of various learners. Ensemble models can be used in the estimation of carbon footprint to incorporate, in their work, the energy consumption, distance in transportation, manufacturing processes, and consumer behaviours in order to produce more credible estimates of emissions.

Fig 1. depicts the general structure of the carbon footprint estimation system, in which user, product and shipping data is processed and analyzed through ensemble learning models to produce emission reports and sustainability insights.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 14, Issue 1, 2026

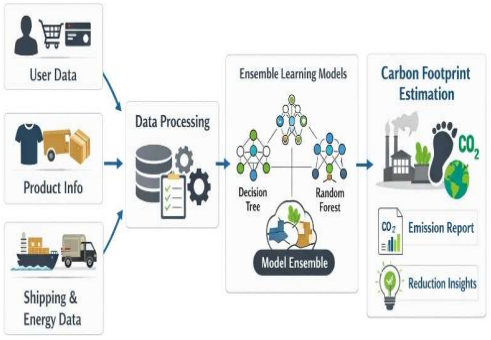

Fig 1.Carbon Footprint estimation System

The proposed framework is systematic in its pipeline since it starts with the data collection by various means including smart meters, transaction logs, and sensors of the environment. This data is taken through preprocessing processes such as noise elimination, normalization, and treatment of missing values. The features are then identified to capture the relevant features that will be used in representing emission-related factors. Those are the features that are used as inputs to several learning algorithms, the results of which are used together to create final emission estimates. The architecture allows scalability and flexibility to various application areas because of its modular architecture.

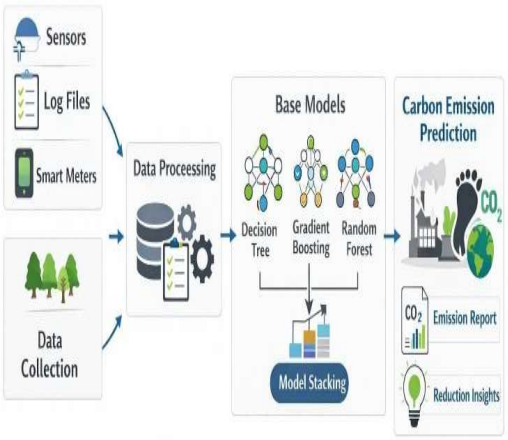

Fig 2.Emission Prediction Ensemble Learning Architecture

Literature Survey

Table 1. Overview of Correlated Literature on Carbon Footprint Estimation with machine learning.

Ref. No.	Carbon Emission Domain Addressed	Algorithm Used	Performance Metrics	Accuracy / Performance	Limitations
[1]	Household energy emission prediction	Linear Regression	MAE, RMSE, R^2	$R^2 \approx 0.75–0.82$	Cannot model nonlinear relationships; low adaptability
[2]	Transportation carbon emission estimation	Support Vector Machine (SVM)	RMSE, MSE	Accuracy ≈ 80–88%	Sensitive to kernel choice and parameter tuning
[3]	Industrial emission classification	Decision Tree	Accuracy, MAE	Accuracy ≈ 78–85%	Prone to overfitting; unstable with noisy data
[4]	Smart grid energy emission forecasting	Random Forest	RMSE, R^2	$R^2 \approx 0.85–0.90$	High computational cost; large memory usage
[5]	Manufacturing process emission prediction	Gradient Boosting (GBM)	MAE, RMSE	Error reduction ≈ 15–20%	Long training time; sensitive to hyperparameters
[6]	Industrial energy consumption analysis	Artificial Neural Network (ANN)	MSE, R^2	$R^2 \approx 0.88–0.92$	Requires large datasets; high training complexity
[7]	Vehicle emission prediction	XGBoost	RMSE, MAE	Accuracy ≈ 90–93%	Overfitting risk; requires regularization

The application of machine learning and ensemble methods of learning has been used in other areas of carbon emission to make predictions more accurate and dependable. Table 1 does a summary of the existing works in terms of algorithms, performance metrics and limitations.

Methodology

Data Processing

The proposed system has a systematic pipeline comprising of data collection, preprocessing, feature engineering, model training, and prediction. The data is gathered using various sources and include records of energy consumption, transportation data, industrial production data, and e-commerce transaction data. These data sets are some of the major emission-related variables like the fuel consumption, electricity consumption, distance covered, and volume of production.

Preprocessing includes processing of the missing values, deletion of outliers and normalization of features through min-max scaling is shown in equation (1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where (X) is the initial feature value, and (X_{min}), (X_{max}) are the minimum and maximum value, respectively.

The feature engineering is carried out to derive meaningful features like average energy usage, intensity of emissions, and frequency of transaction. Correlation analysis and thresholding of the variance is used to drop redundant features and dimensionality reduction.

Machine Learning Models

The model uses a variety of base learners to identify a variety of data features.

- Linear Regression (LR)

Linear Regression is a model that is used to approximate the relationship between a dependent variable and several independent variables as shown in equation (2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (2)$$

Where,y = predicted output variable, β_0= intercept term, $\beta_1, \beta_2, \ldots, \beta_n$= regression coefficients, $x_1, x_2, \ldots, x_n$= input features.

- Support Vector Machine (SVM)

Support Vector Regression predicts the result with the help of an optimal hyperplane as shown in equation (3):

$$f(x) = w^T x + b \quad (3)$$

Where, $f(x)$= predicted output, w= weight vector, x= input feature vector, b= bias term, $w^T x$= dot product of weights and inputs.

- Random Forest (RF)

Random Forest is a set of several decision trees that are used to enhance the precision of predictions as shown in equation (4):

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad (4)$$

Where, $\hat{y}$= final predicted value, N= number of trees, $T_i(x)$= prediction of the i^{th}tree, x= input feature vector.

- Gradient Boosting (GBM)

Gradient Boosting gradually constructs models so as to reduce the error in prediction as shown in equation (5):

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x) \quad (5)$$

Where,$F_m(x)$ = prediction at iteration m, $F_{m-1}(x)$= previous model output, α= learning rate, $h_m(x)$= weak learner at step m, m= boosting iteration index.

Ensemble Learning Framework

In order to improve the predictive power, stacking-based ensemble approach is embraced. Independent training of base models (LR, SVM, DT, RF, GBM, ANN) is done on the preprocessed dataset. They are run individually and the results are taken as meta features into a meta-learner, which is usually a linear regressor or gradient boosting model.

In equation (6), where $M_i(x)$ is the prediction of the i^{th} base model. The group performance is calculated as:

$$\hat{y}_{ensemble} = \sum_{i=1}^{k} w_i M_i(x) \quad (6)$$

where w_i refers to the weight that is optimized on each learner.

Cross-validation is also used to prevent the overfitting a model and unbiased generation of meta-features. The grid search is used to optimize hyperparameters.

Model Evaluation

The performance of the models is measured on standard measures:

- Mean Absolute Error (MAE):

MAE is the mean size of errors in prediction and it is calculated as shown in equation (7):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (7)$$

Where, n= number of samples, y_i= actual value of the i^{th}sample, $\hat{y}_i$= predicted value of the i^{th}sample, $|\cdot|$= absolute value.

- Root Mean Square Error (RMSE):

RMSE is more negative to big errors and it is calculated as shown in equation (8):

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 14, Issue 1, 2026

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

Where, n= number of samples, y_i= actual value, $\hat{y}_i$= predicted value.

- Coefficient of Determination (R^2):
R^2 shows the goodness of fit of the model and it is calculated as shown in equation (9):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (9)$$

Where, y_i= actual value, $\hat{y}_i$= predicted value, $\bar{y}$= mean of actual values, n= number of samples.

These measures have high fidelity of comparison between individual and ensemble models.

Results and Discussion

This is the experimental evaluation of the proposed carbon footprint estimation framework using multiple machine learning models and a stacking-based ensemble approach. The models were evaluated using MAE, RMSE, R^2.

Table 2. Performance Evaluation Metrics and Their Mathematical Equations

Metrics	Equation
Accuracy	$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}}$
Recall	$\text{Recall} = \frac{\text{TP}}{\text{FP} + \text{FP}}$
Precision	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
F1-Score	$\text{F1 score} = \frac{(2 * (\text{Recall} * \text{Precision}))}{(\text{Recall} + \text{Precision})}$

Table 2 depicts the effectiveness of the proposed carbon emission prediction model is evaluated with the help of performance evaluation metrics. Accuracy is used to measure the level of overall correctness of predictions, Precision and Recall are used to determine the capacity of the model to recognize positive cases. The F1-Score is used to measure the performance of a model in a balanced manner using Precision and Recall. These measures are based on the confusion matrix and guarantee the sound and stable assessment of the classification system.

Table 3. Performance Comparison of ML Models for Carbon Footprint Estimation

Model	Precision	Recall	F1-Score	RMSE
Linear Regression	0.81	0.78	0.79	0.46
Support Vector Machine	0.85	0.82	0.83	0.38
Decision Tree	0.87	0.84	0.85	0.34
Random Forest	0.92	0.90	0.91	0.24
Gradient Boosting	0.93	0.91	0.92	0.21

Table 3 provides a performance comparison of various machine learning models based on carbon footprint estimation on Precision, Recall, F1-Score, and RMSE as assessment metrics. Conventional models like Linear Regression and Support Vector Machine perform moderately with comparatively large values of errors. Decision Tree and Random Forest are tree-based models, which have better accuracy and lower RMSE. Gradient Boosting also increases the prediction performance. The Stacking Ensemble model has the best scores in the Precision, Recall, and F1-Score and the lowest RMSE value, which means that it is more effective and reliable in estimating the carbon footprint.

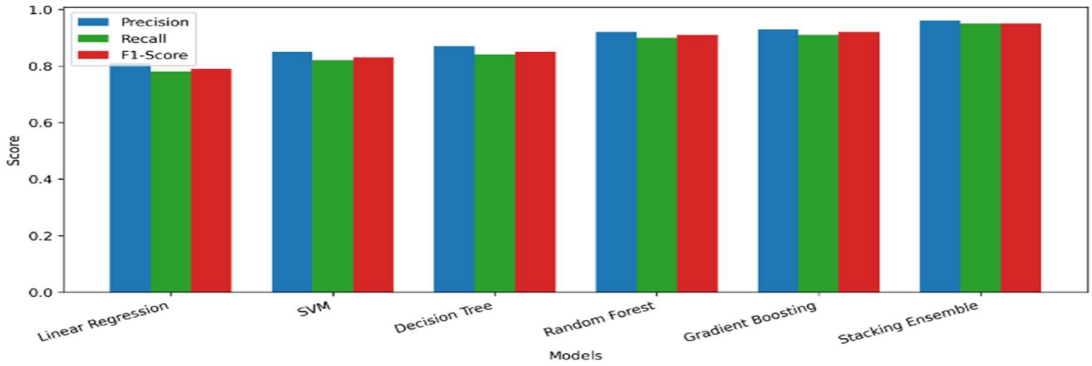

Fig 3. Comparison of Precision, Recall, and F1-Score of Machine Learning Models for Carbon Footprint Estimation

Fig 3 shows the comparison between Precision, Recall, and F1-Score of various machine learning models that are applicable in estimating carbon footprint. Linear Regression and Support Vector Machine have moderate performances whereas tree-based models like Decision Tree and Random Forest are better. Gradient Boosting also enhances the accuracy of the prediction. The Stacking Ensemble model suggested reveals the greatest Precision, Recall, and F1-Score, which means that it has a better classification and reliability in estimating carbon emissions.

Conclusion

The paper discussed an ensemble learning-based model of the correct estimation of carbon footprint with the help of several machine learning algorithms. The suggested methodology also combines data preprocessing, feature engineering, and various predictive models such as Linear Regression, the Support Vector Machine, the Random Forest, the Gradient Boosting, the Artificial Neural Network and a stacking ensemble approach. As evidence of the expansive advancement in enhancing the prediction performance, it is established through experimental results that ensemble learning can outperform individual models efficiently in lowering the MAE and RMSE values besides having a significant coefficient of determination. The stacking ensemble method is effective in bias and variance reduction and hence leads to better generalization and robustness. Moreover, emission levels have been properly categorized through classification analysis by a confusion matrix. The results suggest that the carbon emission monitoring can be efficiently and successfully approached with the help of an ensemble-based model. The suggested framework may be implemented in smart cities, industrial systems, and digital commerce platforms to facilitate the process of making data-driven decisions and sustainable development.

References

1. Chen, Y. M., Chen, Q., Liu, F. G., Zhou, Q., Jin, Q., & Lin, X. (2025). Research Progress on Carbon Emissions from Household Energy Consumption: a Global Perspective. Applied Ecology and Environmental Research, 23(6), 11311-11330.
2. Liu, Z., & Qiu, Z. (2023). A systematic review of transportation carbon emissions based on CiteSpace. Environmental Science and Pollution Research, 30(19), 54362-54384.
3. Rahman, M. M., Shafiullah, M., Alam, M. S., Rahman, M. S., Alsanad, M. A., Islam, M. M., ... & Rahman, S. M. (2023). Decision tree-based ensemble model for predicting national greenhouse gas emissions in Saudi Arabia. Applied Sciences, 13(6), 3832.
4. Hassan, M. A., Salem, H., Bailek, N., & Kisi, O. (2023). Random forest ensemble-based predictions of on-road vehicular emissions and fuel consumption in developing urban areas. Sustainability, 15(2), 1503.
5. Ojadi, J. O., Onukwulu, E., Odionu, C., & Owulade, O. (2023). AI-driven predictive analytics for carbon emission reduction in industrial manufacturing: a machine learning approach to sustainable production. International Journal of Multidisciplinary Research and Growth Evaluation, 4(1), 948-960.
6. Feng, W., Chen, T., Li, L., Zhang, L., Deng, B., Liu, W., ... & Cai, D. (2024). Application of Neural Networks on Carbon Emission Prediction: A Systematic Review and Comparison. Energies, 17(7), 1628.
7. Zhang, L., Lu, G., Yan, X., Xia, P., Chen, Z., & Wu, D. (2025). A differential evolution optimized hybrid XGBoost for accurate carbon emission prediction. Environmental Modelling & Software, 106627.
8. Wang, M., Yu, J., Zhou, M., & Cao, W. (2026). Analysis and prediction of carbon emissions from public building operation based on stacking ensemble strategy: the case of Xi'an. Energy, 139890.
9. S. Kondapalli, M. Dudala, K. K. Kumar, K. Spurthi, K. S. Kumar and R. D. Kumar, "Deep Learning

Convolutional Nets: Intelligent System for Paddy Leaf Disease Diagnosis," 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON), Tumkur, India, 2025, pp. 1-8, doi: 10.1109/SSITCON66133.2025.11341943.

10. Y. Sowjanya, S. Gopalakrishnan and R. D. Kumar, "Internet of Things in Health Care: Motivation and Challenges: A Survey," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10725769.
11. Vojja, L., Kumar, R.D., Sivaprasad, P.V.S., Satyanarayana, A. (2025). Encryption with Identity Based Approach for Versatile Encrypted Data Sharing in Public Cloud. In: Farhaoui, Y., Herawan, T., Lucky Imoize, A., Allaoui, A.E. (eds) Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment. ICAISE 2024. Lecture Notes in Networks and Systems, vol 1397. Springer, Cham. https://doi.org/10.1007/978-3-031-90921-4_81