Journal of Information Systems Engineering and Management
2025,10(5s)
e-ISSN:2468-4376
https://www.jisem-journal.com/

Research Article

Predicting Soil Microbial Biomass Carbon Using Stacking Machine Learning Techniques to Enhance Soil Health

Phalguna Siddhartha Reddy Chilukuri[1], Narra Snehith[2], Aaskaran Bishnoi[3], Dr. Gurwinder Singh[4], Dr. Aman Kaushik[5]

[1]*Apex Institute of Technology (AIT-CSE), Chandigarh University, Gharuan , Mohali, Punjab*
phalgunasiddharthareddy@gmail.com
Orcid:- 0009-0006-5975-9048
[2]*Apex Institute of Technology (AIT-CSE)*
Chandigarh University, Gharuan , Mohali, Punjab
narrasnehith10@gmail.com
[3]*Assistant Professor*
Apex Institute of Technology (AIT-CSE), Chandigarh University, Gharuan, Mohali, Punjab
aaskaran29@gmail.com
[4]*Apex Institute of Technology (AIT-CSE), Chandigarh University, Gharuan, Mohali, Punjab*
gurwinder.e11253@cumail.in
[5]*Associate Professor, Apex Institute of Technology (AIT-CSE), Chandigarh University, Gharuan, Mohali, Punjab*
Amankaushik19ycs1029@gmail.com

ARTICLEINFO

Received: 08 Oct 2024
Revised: 09 Dec 2024
Accepted: 24 Dec 2024

ABSTRACT

Soil microbial biomass carbon (SMBC) is an important factor that affects soil fertility ,biogeochemical cycling and also carbon elimination from atmosphere .Traditionally, the determination of SMBC has been labor - intensive and complex, relying on methods such as the chloroform fumigation-extraction technique, which are both time-consuming and error-prone. The most recent breakthroughs in artificial intelligence have revealed a promising application of the use of AI for the automation and enhancement of the precision of SMBC estimation, especially by the application of deep learning models, such as ANN. Yet, the development of AI applications in this field is relatively underdeveloped and less explored in terms of its practical application and performance. This paper presents an innovative approach for SMBC values using machine learning techniques for improved precision. We use a stacking method, which utilizes advantages of a two machine learning models (lightgbm for numerical features and catboost for categorical features) with a meta-learner , such as Random Forest. The results are promising, as our method yields an R-squared value of 0.75 and an MSE of 0.23, thus making it a useful tool for SMBC estimation. Such approach shows significant steps forward on the challenges faced by other classical approaches; thus, they will represent more efficient, reliable alternatives in large scale assessments for soil health along with environmental monitoring.

Keywords: Soil Microbial, Machine Learning.

INTRODUCTION:

Soil is vital in maintaining terrestrial ecosystems; it influences plant growth, regulates water cycles, and contributes significantly to the global carbon cycle. Soil health and fertility are crucial for agricultural sustainability, influencing crop yields and environmental resilience. Among the many indicators of soil health, SMBC has emerged as an important parameter because it directly reflects the biological activity in soil as well as nutrient flow and general maintenance of the soil structure [14]. The monitoring of SMBC provides an insight into the soil's capability to ensure growth of plant and habitat functions, and therefore, it is one of the key metrics for precision agriculture and environmental management.

SMBC is the measure of carbon in soil microorganisms, which is a labile fraction of soil organic matter. This biomass is crucial for better nutrient cycling and also how well organic matter is decomposed, and stability of soil aggregates. The abundance and diversity of microbial communities have a direct impact on crucial soil functions such as nutrient

availability, carbon sequestration, and soil fertility, all of which are necessary for maintaining agricultural productivity and environmental stability [15]. Studies have repeatedly shown that greater amounts of SMBC are associated with better soil structure, greater nutrient availability, and higher soil carbon storage, leading to improved crop yields and long-term soil health [16].

In addition, microbial biomass indicates the changes in the soil conditions and responds rapidly to changes in land use, tillage, or fertilization. By this way SBMC can be a useful tool to assess how different farming techniques effect soil health and to guide sustainable land management decisions [17]. Moreover, SMBC has an important role in decreasing climate change by enhancing soil carbon sequestration. Since soils account for the largest carbon sink on land, understanding microbial processes that regulate carbon storage can help reduce carbon emissions [18].

Healthy soil microbial communities are essential in a world where global challenges face agricultural systems, such as food security and climate change. Therefore, accurate measurement and prediction of SMBC are very vital for effective soil management and maximization of agricultural productivity. Traditional methods for measuring SMBC, such as fumigation-extraction and fumigation-incubation, have been used long enough but are usually labour-intensive, costly, and time-consuming, hence impossible to scale up or applicable for large-scale or even continuous monitoring [19]. These limitations imply that there is a dire need for more efficient, scalable, and cost-effective techniques for monitoring SMBC, which could offer real-time, large-scale assessments of soil.

Recent development in AI and ML has thus been promising in the automatization of SMBC estimation as a potential solution to traditional methods' challenges. Because AI and ML showed promising results in other fileds like efficient waste processing which can affect health of the soil[23]. AI-based methods have the potential to offer even more accurate, efficient, and scalable measurements of the soil microbial biomass, thus leading to better monitoring and management of soil health. Based on the evidence that mounts daily, the importance of SMBC in agriculture and environmental sustainability has been acknowledged [14], and this calls for developing advanced AI models to predict SMBC for the betterment of precision agriculture and soil management strategies.

LITERATURE REVIEW:

Historically , Researchers have proposed different methods to find the Soil Microbial Biomass Carbon (SMBC). Traditional, these methods could be broadly classified into two general types: physiological approaches - such as Chloroform Fumigation Incubation Method and Substrate-Induced Respiration Method, and chemical techniques - encompassing Chloroform Fumigation Extraction Method and ATP Determinations [1]. Additionally, there are other methods also like Extraction method to find SMBC [2]. Moreover, It can be also measured by Substrate-Induced Respiration method [3].However, Despite their widespread application,they suffer from significant limtations However, machine learning techniques do offer more efficient and accurate solutions.

Recent advances in ML have released new approaches for predicting soil microbial biomass carbon (SMBC), which is now feasible to explore complex, non-linear interactions among soil attributes and microbial biomass. For example, Pellegrini et al. successfully demonstrate the utility of using ANNs in predicting SMBC in vineyard soils, with improved accuracy as compared to conventional regression based methods. Their research revealed that SOM was the most significant parameter for predicting SMB-C, outperforming linear regression models mainly for low and mid-range values of SMB-C [8]. In a similar study, ML techniques such as Random Forest and AdaBoost were applied to correlate soil microbial biomass with carbon sequestration in soil in agroecosystems. Indeed, it was clear that such ensemble models outperformed classical methods in predicting SMBC and identified labile carbon and above-ground biomass as the drivers of microbial biomass with various soil-health management practices. Opportunities for optimizing agricultural management to improve soil health and associated microbial biomass existed as presented in the paper [7]. The application of machine learning techniques combined with sensor fusion, including both UAV imagery and terrestrial sensor information, improved the soil organic matter (SOM) prediction significantly. The Random Forest model proved to be the most accurate model for this purpose (RMSE = 0.13 and R^2 = 0.68). These developments have shown that machine learning can solve the shortcomings of traditional approaches, improving the accuracy and efficiency of predictions about soil microbial biomass and organic matter [9].

Further to these inventions, studies have been undertaken on the application of ML models to analyze the SMBC global trends and their sensitivity to environmental factors. An example study used Random Forest models to analyze the SMBC global trends between 1992–2013 and found a global SMBC decrease of 3.4% over this period, which was largely controlled by temperature increases and in northern regions. This study used a dataset with SBMC and also

integrated with environmental data layers, which identified a great need for an enhanced understanding of the impacts of climate change on microbial biomass in soil ecosystems [4]. Another study focused on dryland agriculture, where Artificial Neural Networks (ANNs) were used to predict soil quality indices based on different physical, chemical, and fertility parameters. The ANN models provided an excellent accuracy in classifying soils into five different quality levels, which may have significant implications for soil health management in arid areas with a highly impressive R^2 value of 0.97–0.98 [5]. Ensemble learning techniques have also been effective in determining the interaction between SMB-C and soil carbon sequestration through various studies concerning the effect of land use, seasonality, and depth of soil. These models, especially Random Forest and AdaBoost, performed better compared to traditional approaches and hence better explained the optimization of soil health management practices associated with sustainable agriculture [7]. Stacking ensemble models such as XGBoost, LightGBM, and CatBoost have enabled downscaling of soil moisture data into finer spatial resolutions for highly accurate predictions of soil moisture, which is crucial in the improvement of soil moisture management and irrigation methods [13].

Despite these developments, the application of ML to SMBC prediction is still quite nascent. Most of the studies focus on a specific dataset like vineyard soils [8] or broader trends, such as global SMBC changes [4], while there has been little research into SMBC prediction over diverse soil types and under different environmental conditions. In addition, deep learning architectures, such as artificial neural networks (ANNs), have been more visibly applied; on the contrary, machine learning methodologies like Random Forest, XGBoost, and Support Vector Machines (SVMs), along with ensemble approaches, have been relatively rarely used in this domain. This situation presents a clear gap in research, because machine learning models are usually less resource-intensive and are capable of providing similar or even better predictive performance if fine-tuned appropriately. This study addresses the gap with machine learning methods, proposing a new method using stacking approach to predict SMBC.

MATERIALS REQUIRED:-

I. Dataset exploration

The Dataset we used for our model was provided by the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)[24]. This dataset includes 3,422 datapoints from 315 research papers published between 1970s and 2012[25].

Below figure shows at which places the data was collects and number of data points and also what was the biome type

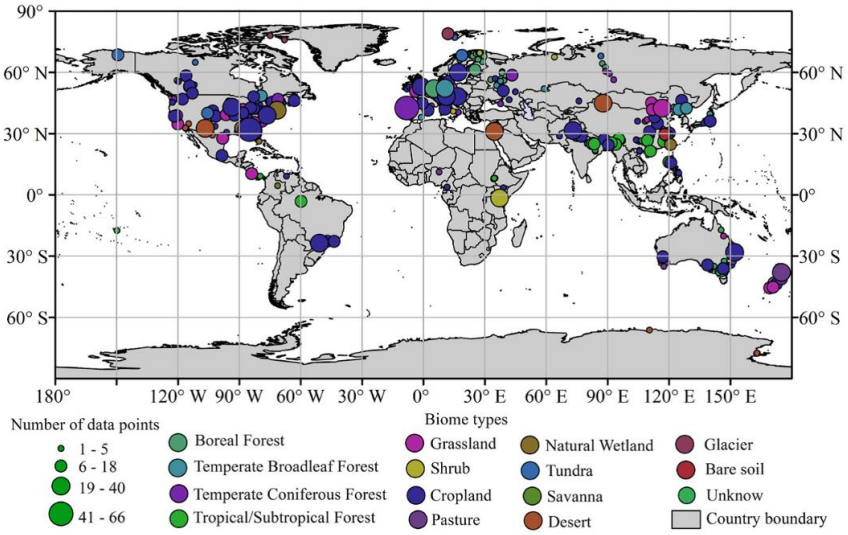


Figure A:- locations of the datapoints that were used for this study with geographical coordinates[25]

Features of the dataset

1.it gives the concentration of soil microbial biomass carbon (C),total nitrogen , soil organic carbon and total phosphorus at biome and global scales

2.this data was complied from soil samples some from at 0-15cm depth or from 0-30cm depth.

3.they also listed latitude, longitude for majority of data points and additionality they also provided extra soil properties , climate data ,site characters such as pH value ,mean temperature which are useful for our model

4.Data was Collected all around the world which gives massive diversity and advantage from training our model

Table shows statistical description of numerical attributes in the data set

Column Name	count	mean	std	min	25%	50%	75%	max	Missing Values
Latitude	2705	42.66	12.52	0.1	32.6	42.95	51.8	79	162
Longitude	2109	69.77	59.8	0.37	11.6	37.62	119	177.9	164
Elevation	655	575.1	566.8	0	145	430	785	2400	2632
MAT	1056	13.71	7.772	0	7.3	12.7	19.1	30	2295
MAP	1274	901.4	720.4	0.1	400	750	1300	5100	2138
Soil_microbial_biomass_carbon	3183	62.37	128.4	0.04	14.5	27.17	57.25	1508	240
Soil_microbial_biomass_nitrogen	1440	9.388	15.09	0.03	1.95	4	9.703	126.4	1983
Soil_microbial_biomass_phosphorus	707	2.178	3.738	0.02	0.36	0.65	2.372	48.45	2716
Soil_organic_carbon	2946	3883	7426	8.33	900	1442	3440	63917	477
Total_nitrogen	2162	244.9	320.2	4.93	78.6	128.6	271.4	3071	1261
Total_organic_phosphorus	535	24.43	29.32	0.41	10.5	18.06	29.16	271	2888
pH	2161	6.024	1.76	0.8	5	6	7	63.86	1246
Date	219	1999	8.237	1981	1993	1999	2005	2010	1039
Upper_depth	2961	1.586	4.197	0	0	0	0	50	385
Lower_depth	2949	13.33	6.776	0	10	10	15	60	473
Depth	12	1.083	2.353	0	0	0	1	8	385

II.ML Models Used:-

2.1LightGBM

A leaf-wise technique was used in LightGBM also it uses some advanced techniques like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for handling high dimensional data efficiently [20]. It, therefore, trains faster, makes more accurate predictions using lesser memory. GOSS mainly focuses on gradients whereas EFB tries to reduce feature dimensionality by using sparse features [20]

Mathematically, objective function can be minimized by:

$$L(\theta) = \sum_{i=1}^{n} l\big(y_i, f(x_i;\theta)\big) + \Omega(f)$$

Here , l as loss function , $f(x_i,\theta)$ as predictions , $\Omega(f)$ as regularization term

Below flowchart It illustrates the workflow of LightGBM, showing histogram-based feature processing in histogram Algorithm, leaf-wise learning, and iterative gradient boosting to optimize residuals in constructing an ensemble model.

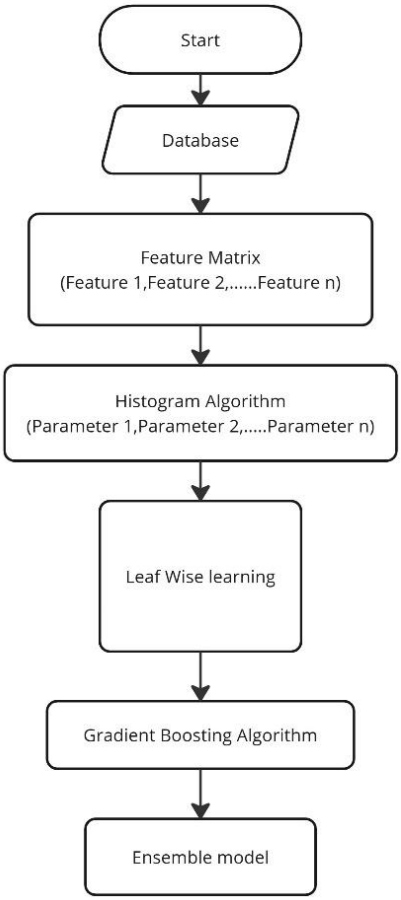


Figure:-Workflow of LightGBM

2.2Catboost

CatBoost specializes in categorical data using ordered boosting with an approach to overfitting via symmetric trees, hence robust performance on the high cardinality categorical feature sets[24]. Ordered boosting ensures against data leakage by training over permutations of the dataset.[21]

Mathematically , The loss function of catboost can be written as

$$L(\theta) = \sum_{i=1}^{n} l\big(y_i, f(x_i;\theta)\big) + \lambda\|\theta\|^2$$

Here, l is loss function and $\lambda||\theta||^2$ Is regularization term which regulates overfitting

The figure illustrates CatBoost's iterative training process, in which features are synthesized, and misclassified samples are assigned greater weights in subsequent iterations. The ultimate prediction constitutes a weighted average of all iterations, thereby ensuring enhanced accuracy.

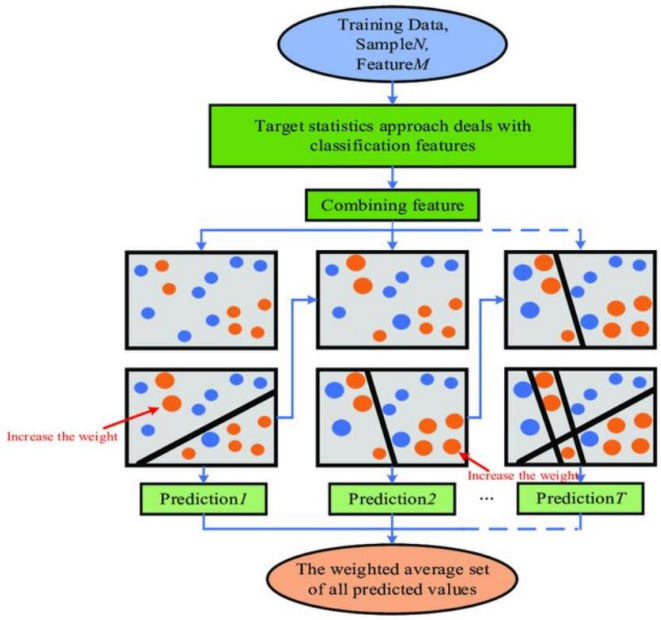


Figure:-Workflow of CatBoost

2.3Stacking and ensemble method

Stacking is a ensemble method where it combines predictions from a variety of base models with a meta-model for final predictions, making use of their complementary strengths[22]. This improves overall accuracy and robustness by learning higher-level patterns[22]. The main use of stacking it exploits the strength of base models while removing the individual weakness through meta-learning

Mathematically , Working of stacking can be shown as

$$\hat{y} = g(h_1(x), h_2(x), \dots, h_n(x))$$

Here $h_i(x)$ are predictions of base models , g is meta-model that is trained on these predictions

Now, Meta-model minimizes by

$$L(\theta) = \sum_{i=1}^{n} l\big(y_i, g(h(x_i); \theta)\big)$$

This ensures optimal integration of diverse models

The process of stacking ensemble learning can be repseresent by in a figure 1 . Here, At level-0, several base learners (L1, L2, ., LT) are trained independently on the dataset to produce predictions. These generated predictions, in conjunction with the actual classification outcomes, constitute new data for level-1, wherein a meta-learner amalgamates them to arrive at the final prediction. This hierarchical structure improves model diversity and accuracy by utilizing the advantages of each individual learner.

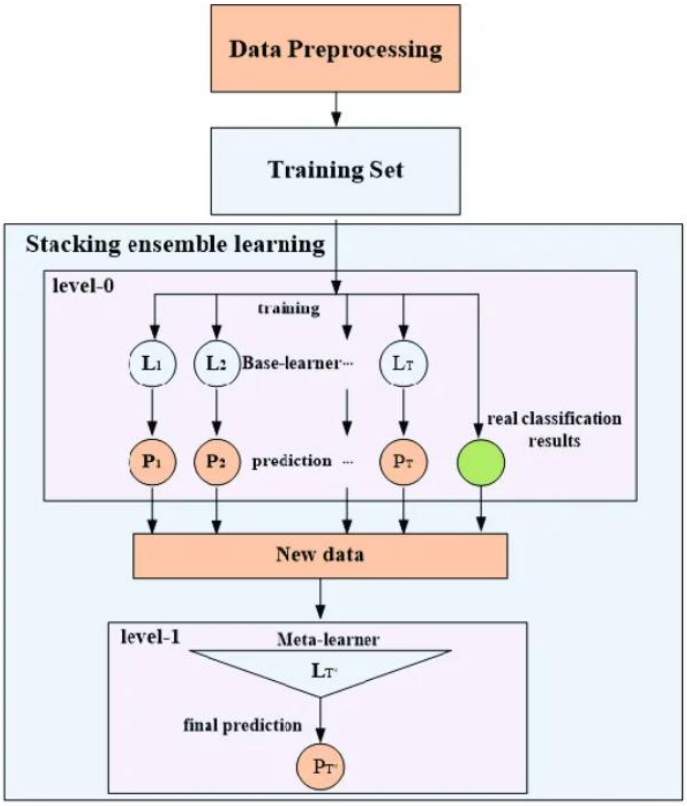


Figure:-Stacking ensemble learning process

2.4Optuna Hyperparameter optimizer

Optuna Hyperparametertuning optimization technique Optimizes objective functions by using algorithms like Tree-structured Parzen Estimators It accelerates the model's optimization with lesser iteration It ensures that it makes exploration of hyperparameters very efficient, thereby boosting model accuracy and reducing computation overhead TPE constructs probability distributions to focus on promising parameter ranges for faster convergence [24].

Mathematically,

Optuna minimizes an objective :

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

Where θ is the parameter space and f(θ) is validation loss

III. Evaluation metrics

We have used two evulation metrics to determine the result of our model

A.R2

It measures the fit of our regression model to the information. It is an indicator of the proportion that could be predicted in the dependent variable based on using independent variables. An R^2 value of 1 tells that model would exactly explains all variances in the predicting variables. On the other end, an R^2 with a value of 0 would mean that the variation could not be explained at any given point by the model constructed.

Mathematically,

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

Here ,

y_i -actual value

$\hat{y}_i$ – predicted value got from model

$\bar{y}$ -mean of actual values

n is number of data points

B.MSE

An average of squared differences comprised actual and predicted values is known as MSE. It gives a sense of how much a model's prediction differs from the true value. In other words,The lesser the MSE, the greater the fit of a model.

Mathematically,

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

Where,

y_i actual value

$\hat{y}_i$ predicted values

n number of data points

Methodology: -

I. Data Collection

Loaded the dataset “Soil_Microbial_Biomass_C_N_P” using pandas

II. Data Preprocessed

We have preprocessed the given dataset so that it can be useful for our model. We have taken Preprocessing techniques like cleaning the dataset (metadata rows were removed and resetting the index, Converting Specific columns to numeric datatypes, handling missing values in target variable). We also defined features by looking their relations with target variable. Numerical Features are Latitude, Longitude, Soil organic carbon, Total nitrogen, pH, Mean Annual Temperature (MAT), Mean Annual Pressure (MAP) and categorical features are Biome, Vegetation Then target variable was selected from the dataset is Soil Microbial Biomass Carbon. Moreover, log transformation was applied on this target variable because its range is very large and that can affect model’s performance. In next step features were stated as X and target y. Finally Numerical Features were even preprocessed like imputing missing values, applying standard scaling and categorical Features were also preprocessed like handling values that are missing and adding one-hot encoding as transformation.

III. Splitting Data

This preprocessed data was split into training 80% and testing 20% Sets

IV. Model Definition

We used Stacking Ensemble method where LightGBM (which is best for numerical data) and Catboost(which is best for categorical data) as base model and random regressor is used as Meta Learner or final Estimator. By this way we have leveraged only advantages of both model as we used both numerical and categorical features.

V. Model Training

Both Models were trained using training dataset by trying different set of hyperparameters of both models. This process was conducted by using Optuna. After using choosing hyperparameters for both models which are optimal for our training dataset.

The models were later retrained using those optimal hyperparameters on entire training set to build final stacking model where the meta learner decides the result from those base models

VI. Model Evaluation

After the model was trained, it was tested using test data. We used R2 and MSE as our Evaluation metrics and Graphs are also used for Visual understanding

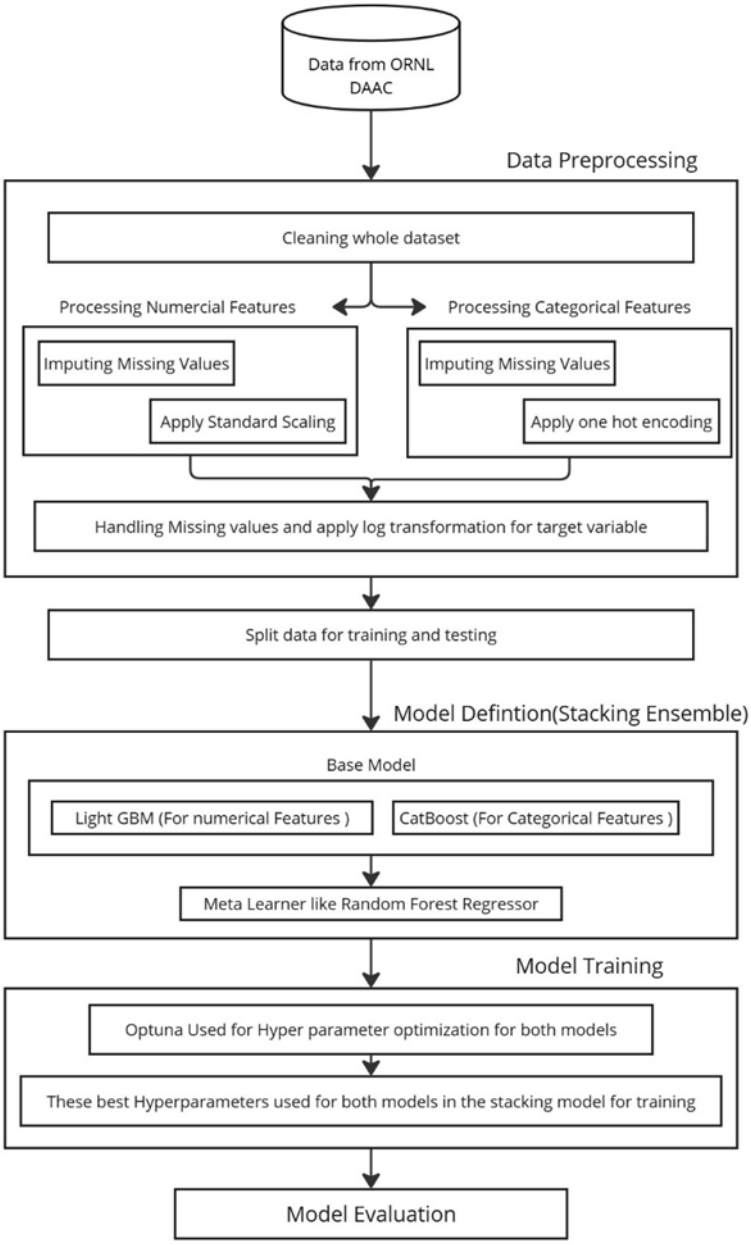
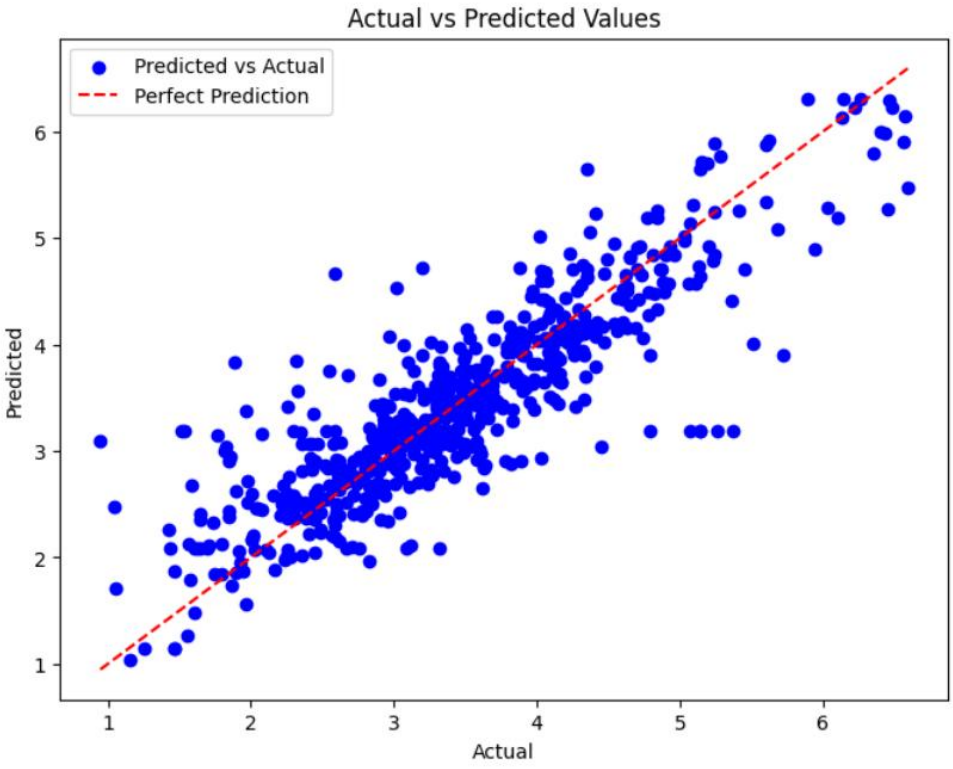


Figure :- Flowchart of the Methodology

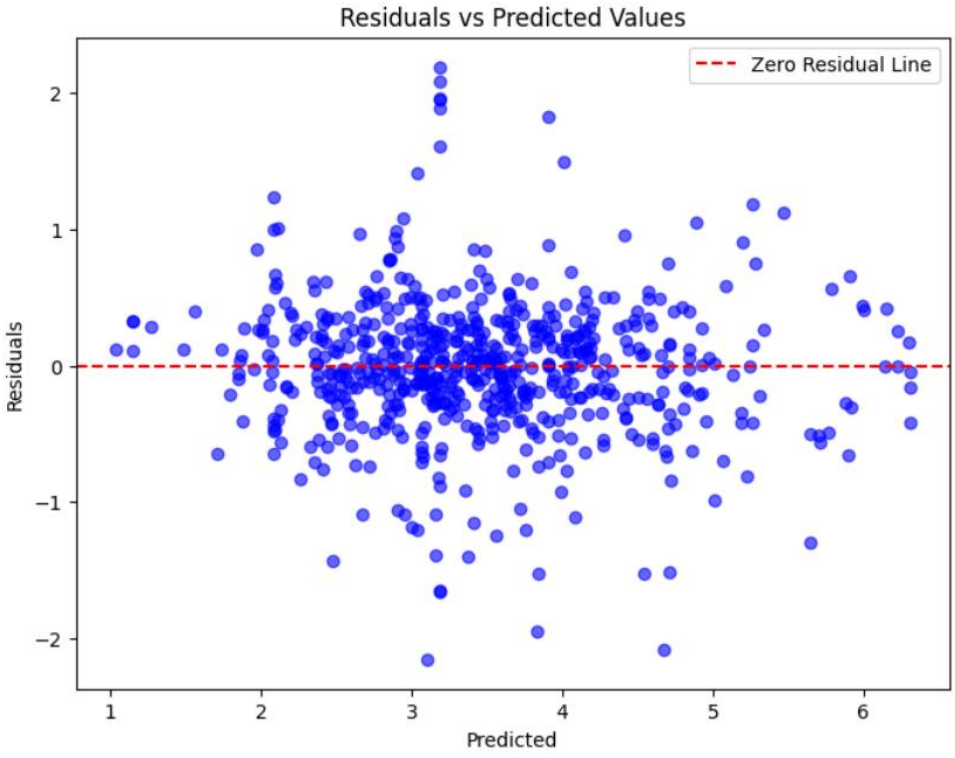
RESULTS

We have evaluated model's performance using two main metrics one is R2 and another is MSE.This model achieved R2 value of 0.755 and MSE is 0.231.

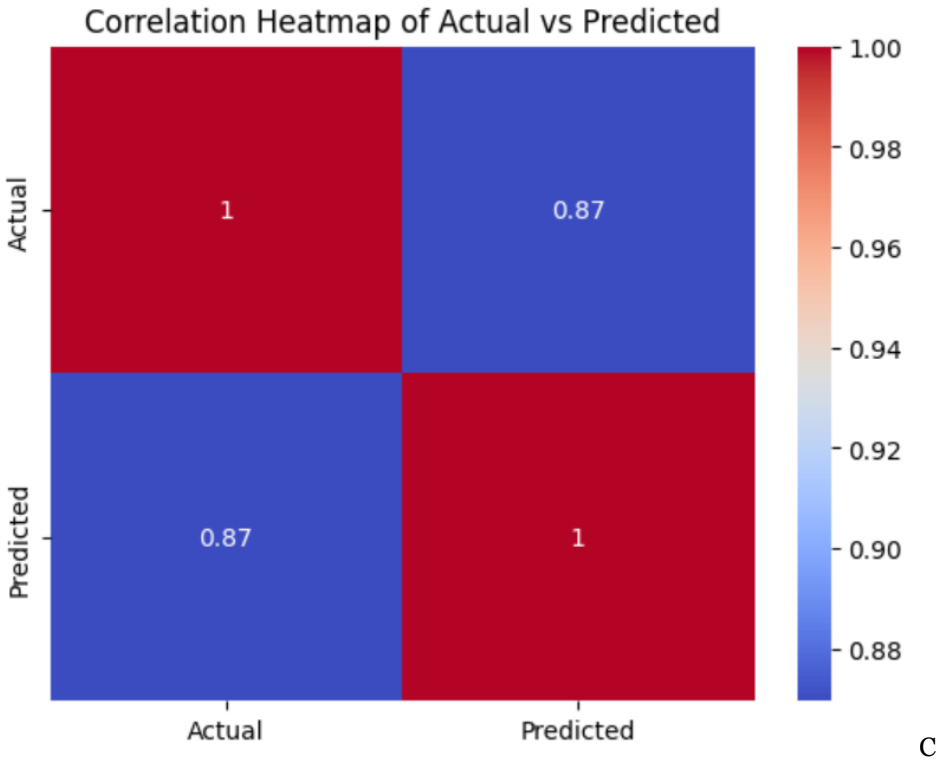
To interpret the model performance even clearly, The following visualization can help to interpret model's performance even more. Figure A Shows a plot between Predicted values and actual values and in that red dotted line shoes prefect prediction , Graph **B** shows Residuals V/s Predicted values and Finally Figure **C** Shows Correlation Heatmap between Actual v/s Predicted


Actual vs Predicted Values
Predicted vs Actual
Perfect Prediction
Predicted
Actual
1
2
3
4
5
6

A


Residuals vs Predicted Values
Zero Residual Line
Residuals
Predicted
2
1
0
−1
−2
1
2
3
4
5
6

B



C

CONLUSION AND FUTURE WORK

Soil Microbial Biomass Carbon (SMBC), also known as crucial indicator of soil health and fertility evaluates the biological activities and nutrient conditions of soil. It is tedious on all traditional methods of determining SMBC in comparison to some other soil properties, for instance, pH, which is easy compared. AI advancements in the recent past are tending to change it all. Very minimal studies have been done in applying AI techniques to predicting soil properties, especially SMBC.

In this work, we proposed the new approach on SMBC using the stacked ensemble for prediction. The base learners directly with the Random Forest final estimator component included LightGBM and CatBoost. The combinations of both numerical (like pH, MAT, MAP) and categorical features (Biome, Vegetation) strongly associated with SMBC were derived from data collected from ORNL DAAC. The results summarize model effectiveness, yielding better predictability R^2 0.75 and MSE 0.23.

This work is yet another on applying machine learning to soil science, whereby it becomes easy for both farmers and researchers to estimate SMBC efficiently. As it improves the quality and reduces the costs of precise assessments of soil health, it will further lead to more sustainable agricultural practices in soil management.

FUTURE WORKS:

Though Our Study shows promising results there is still lot of room to improve in this field which leaves much wider scope for more studies and contributions by researchers so that various machine learning and ensemble models could be evaluated and applied to potentially capture more accurate predicted results.

In contrast, datasets will play a vital role in how well prediction models perform. Future studies should add to and develop larger and cleaner databases that complete the schemes in general detail, capturing features of higher variability. They should also include other parameters in environmental, biological, and chemical realms to reflect better the complexity of soil properties.

Thus, further development in this domain might extend the work potential of soil scientists, data scientists, and AI scientists combined. This may well establish a platform for addressing issues concerning the enhancement of efficiency and reliability of models in predicting soil microbial biomass carbon. In the long run, such developments will be aimed at supporting sustainable agriculture and grounds well managing soil health.

REFERENCES

[1] Horwath, W. R. and E. A. Paul. 1994. Microbial biomass. Pages 753-774 in R. W. Weaver, J. S. Angle, P. J. Bottomley, D. F. Bezdicek, M. S. Smith, M. A. Tabatabai, and A. G. Wollum, eds. Methods of Soil Analysis Part 2-Microbiological and Biochemical Properties. Soil Science Society of America, Madison, Wisconsin, USA.

[2] Vance, Eric & Brookes, P. & Jenkinson, D.. (1987). An Extraction Method for Measuring Soil Microbial Biomass C. Soil Biology and Biochemistry. 19. 703-707. 10.1016/0038-0717(87)90052-6.

[3J.P.E. Anderson, K.H. Domsch,A physiological method for the quantitative measurement of microbial biomass in soils,Soil Biology and Biochemistry,Volume 10, Issue 3,1978,Pages 215-221,ISSN 0038-0717.

[4] Patoine, G., Eisenhauer, N., Cesarz, S. *et al.* Drivers and trends of global soil microbial carbon over two decades. *Nat Commun* **13**, 4195 (2022).

[5] El Behairy RA, El Arwash HM, El Baroudy AA, Ibrahim MM, Mohamed ES, Rebouh NY, Shokr MS. An Accurate Approach for Predicting Soil Quality Based on Machine Learning in Drylands. *Agriculture*. 2024; 14(4):627.

[6] Lepcha, N.T., Devi, N.B. Effect of land use, season, and soil depth on soil microbial biomass carbon of Eastern Himalayas. *Ecol Process* **9**, 65 (2020).

[7] Reshmi Sarkar, Anil Somenahally,Machine learning soil-environmental impacts on agroecosystems for relating microbial biomass to soil carbon sequestration,Smart Agricultural Technology,Volume 4,2023,100208,ISSN 2772-3755.

[8] Pellegrini, E., Rovere, N., Zaninotti, S. *et al.* Artificial neural network (ANN) modelling for the estimation of soil microbial biomass in vineyard soils. *Biol Fertil Soils* **57**, 145–151 (2021).

[9] Uddin, M. J., Sherrell, J., Emami, A., & Khaleghian, M. (2024). Application of Artificial Intelligence and Sensor Fusion for Soil Organic Matter Prediction. *Sensors*, *24*(7), 2357.

[10] Nannipieri, P., Ascher, J., Ceccherini, M., Landi, L., Pietramellara, G., & Renella, G. (2017). Microbial diversity and soil functions. *European journal of soil science*, *68*(1), 12-26.

[11] Awais, M., Naqvi, S. M. Z. A., Zhang, H., Li, L., Zhang, W., Awwad, F. A., ... & Hu, J. (2023). AI and machine learning for soil analysis: an assessment of sustainable agricultural practices. *Bioresources and Bioprocessing*, *10*(1), 90.

[12] Awais, M., Naqvi, S.M.Z.A., Zhang, H. *et al.* AI and machine learning for soil analysis: an assessment of sustainable agricultural practices. *Bioresour. Bioprocess.* **10**, 90 (2023).

[13] Xu, J., Su, Q., Li, X., Ma, J., Song, W., Zhang, L., & Su, X. (2024). A Spatial Downscaling Framework for SMAP Soil Moisture Based on Stacking Strategy. *Remote Sensing*, *16*(1), 200.

[14] Bargali S. S. Soil Microbial Biomass: A Crucial Indicator of Soil Health. Curr Agri Res 2024; 12(1).

[15] Carson, J. (2012). *Microbial Biomass Carbon — New South Wales*. SoilQuality.org.au. Retrieved from https://www.soilquality.org.au/factsheets/microbial-biomass-carbon-nsw

[16] Patoine, G., Eisenhauer, N., Cesarz, S. *et al.* Drivers and trends of global soil microbial carbon over two decades. *Nat Commun* **13**, 4195 (2022).

[17] Lepcha, N.T., Devi, N.B. Effect of land use, season, and soil depth on soil microbial biomass carbon of Eastern Himalayas. *Ecol Process* **9**, 65 (2020).

[18] Nguyen, Q. C. (2021). Soil microbial biomass: Growing insight. Eurofins AgroScience.

[19] Hoyle, F., Murphy, D., & Sheppard, J. (n.d.). *Microbial Biomass – Queensland*. SoilQuality.org.au. Retrieved from https://www.soilquality.org.au/factsheets/microbial-biomass-qld

[20]Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

[21] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

[22] Breiman, L. (1996). Stacked regressions. *Machine learning*, *24*, 49-64.

[23]Gude, D. K., Bandari, H., Challa, A. K. R., Tasneem, S., Tasneem, Z., Bhattacharjee, S. B., Lalit, M., Flores, M. A. L., & Goyal, N. (2024). Transforming Urban Sanitation: Enhancing Sustainability through Machine Learning-Driven Waste Processing. *Sustainability*, *16*(17), 7626.

[24] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

[25] Xu, X., P.E. Thornton, and W.M. Post. 2014. A Compilation of Global Soil Microbial Biomass Carbon, Nitrogen, and Phosphorus Data. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.