Research Article

Prediction and Interpretation of Total N and Its Key Drivers in Cultivated Tropical Peat using Machine Learning and Game Theory

Prediksi dan Interpretasi Total N dan Faktor-Faktor Pengontrolnya pada Gambut Tropis yang Dibudidayakan menggunakan Pembelajaran Mesin dan Teori Permainan

Heru Bagus Pulunggono[1]*, Yusuf Azmi Madani[2], Lina Lathifah Nurazizah[3], Moh Zulfajrin[2,4]

[1] Department of Soil Science, Faculty of Agriculture, IPB University, Bogor, 16680, West Java, Indonesia

[2] Bachelor of Agriculture, Department of Soil Science, Faculty of Agriculture, IPB University, Bogor, 16680, West Java, Indonesia

[3] Bachelor of Agriculture, Department of Agronomy and Horticulture, Faculty of Agriculture, IPB University, Bogor, 16680, West Java, Indonesia

[4] Researcher at Soil Chemistry and Fertility Division, Department of Soil Science, Faculty of Agriculture, IPB University, Bogor, 16680, West Java, Indonesia

*email: heruipb@yahoo.co.id

Received: December 2023
Accepted: December 2023
Published: December 2023

p-ISSN: 2723-7974
e-ISSN: 2723-7966
doi: 10.52045/jca.v4i1.592

Website:
https://ojs.untika.ac.id/index.php/faperta

Abstract: Currently, there is a growing interest among research communities in the development of statistical learning-based pedotransfer functions/PtFs to predict mineral soil nutrients; however, similar studies in peatlands are relatively rare. Moreover, extracting meaningful information from these 'black-box' models is crucial, particularly concerning their algorithmic complexity and the non-linear nature of the soil covariate interrelationships. This study employed the Pulunggono (2022a) dataset and the bootstrapping method, to (1) develop and evaluate seven PtF models, including both general linear models (GLM) and machine learning (ML) regressors for estimating total nitrogen (N) in tropical peat that has been drained and cultivated for oil palm (OP) in Riau, Indonesia and (2) explaining model functioning by incorporating Shapley Additive Explanation (SHAP), a tool derived from coalitional game theory. This study demonstrated the superior predictive performance of ML-based PtFs in estimating total N compared to GLM algorithms. The top-performing algorithms for PtF models were identified as GBM, XGB, and Cubist. The SHAP method revealed that sampling depth and organic C were consistently identified as the most important covariates across all models, irrespective of their algorithmic capabilities. Additionally, ML algorithms identified the total Fe, pH, and bulk density (BD) as significant covariates. Local explanations based on Shapley values indicated that the behavior of PtF-based algorithms diverged from their global explanations. This study emphasized the critical role of ML algorithms and game theory in accurately predicting total N in peatlands subjected to drainage and cultivation for OP and explaining their model behavior in relation to soil biogeochemical processes.

Keywords: artificial intelligence, machine learning, pedotransfer functions, Shapley Additive Explanation/SHAP, Shapley value

Abstrak: Saat ini, terdapat minat yang semakin besar di kalangan komunitas peneliti dalam pengembangan fungsi pedotransfer berbasis pembelajaran statistik (statistical learning) untuk memprediksi unsur hara tanah mineral; namun demikian, penelitian serupa di lahan gambut masih jarang dilakukan. Selain itu, mengekstraksi informasi yang berarti dari model 'kotak hitam' ini sangat penting, terutama terkait kompleksitas algoritmik dan sifat hubungan kovariat tanah yang tidak linier. Penelitian ini menggunakan dataset Pulunggono (2022a) dan metode bootstrapping, untuk (1) mengembangkan dan mengevaluasi tujuh model PtF, termasuk model linear umum (GLM) dan regresi machine learning (ML) untuk mengestimasi total nitrogen (N) pada gambut tropis yang telah didrainase dan dibudidayakan untuk kelapa sawit (OP) di Riau, Indonesia, serta (2) menjelaskan fungsi model dengan menggabungkan Shapley Additive Explanation (SHAP), sebuah perangkat yang berasal dari teori permainan koalisi. Studi ini menunjukkan kinerja prediksi yang unggul dari PtF berbasis ML dalam mengestimasi total N dibandingkan dengan algoritma GLM. Algoritma dengan performa terbaik untuk model PtF diidentifikasi sebagai GBM, XGB, dan Cubist. Metode SHAP mengungkapkan bahwa kedalaman

Citation:
Pulunggono HB, Yusuf AM, Nurazizah LL, Zulfajrin M. 2023. Prediction and Interpretation of Total N and its key drivers in cultivated tropical peat using machine learning and game theory. *Celebes Agricultural*. 4(1): 46-68. doi: 10.52045/jca.v4i1.592

pengambilan sampel dan karbon organik (C) secara konsisten diidentifikasi sebagai kovariat yang paling penting di semua model, terlepas dari kemampuan algoritmanya. Selain itu, algoritma ML mengidentifikasi total Fe, pH, dan bobot isi (BD) sebagai kovariat yang signifikan. Penjelasan lokal berdasarkan nilai Shapley mengindikasikan bahwa perilaku algoritma berbasis PtF berbeda dengan penjelasan globalnya. Studi ini menekankan peran penting algoritma ML dan teori permainan dalam memprediksi secara akurat total N di lahan gambut yang mengalami drainase dan budidaya untuk OP dan menjelaskan perilaku model mereka dalam kaitannya dengan proses-proses biogeokimia dalam tanah.

Kata kunci: fungsi pedotransfer, kecerdasan buatan, nilai Shapley, pembelajaran mesin, Shapley Additive Explanation/SHAP

INTRODUCTION

Peatlands are regarded as the most carbon-dense of soil ecosystems. Peat soil primarily had high C content (more than 12%; Soil Survey Staff 2014; Subardja et al. 2014) and was derived from organic matter accumulation under prolonged waterlogging and anaerobic conditions (Noor et al. 2016). Although peatlands only cover about 3% of the total global land surface (Yu et al. 2010), they contain about 644 Gt C or around 21% of the 0-300 cm of total global soil organic C stock (Scharlemann et al. 2014; Leifeld and Menichetti 2018). Furthermore, Indonesian tropical peatlands contain around 13.6 to 40.5 Gt C (Warren et al. 2017) of 152 to 288 Gt C stored by tropical peatlands (Ribeiro et al. 2020), which also plays a significant role in the national agricultural sector. This organic-rich soil developed extensively in the lowland areas of Sumatra and Kalimantan (Anda et al. 2021). It provides a site for agricultural developments, particularly for oil palm plantation/OPP (Ditjen Perkebunan 2011; Koh et al. 2011; Bou Dib et al. 2018).

Nitrogen (N) is one of the major soil macronutrients, which is required at a tremendous amount for sustaining plant growth and development, particularly for OP plants (Engels et al. 2012; Corley and Tinker 2015). N was found to be the second-highest nutrient in mineral or peat soils after C (Marschner and Rengel 2012; Pulunggono 2019: Paleckiene et al. 2021). Therefore, maintaining N on peat soil in OPPs is critical since its deficiency may lead to leaf abnormalities (Broeshart et al. 1957), growth retardation (von Uexkьll and Fairhurst 1999; Mohidin et al. 2015), and low yields (Corley and Tinker 2015). Also, the environmental impact regarding OP-cultivated peatland's capability to emit various greenhouse gasses that can contribute to global climate change with respect to N addition (Melling et al. 2006; Chaddy et al. 2019; Chaddy et al. 2021).

An extensive N collection must be done to attain proper N management at a practical level in OPP. A single block (around 30 hectares) is usually established as the smallest sampling unit for OPP management. Nevertheless, previous research highlighted the spatiotemporal variability of peat nutrients at sub-block scales, along with the gradients of the distance from the oil palm tree, canal and mineral soil border, peat thickness, sampling depth, oil palm age, season, and land use (Pulunggono et al. 2016; Pulunggono 2019; Pulunggono 2020; Pulunggono 2021; Pulunggono 2022a; Pulunggono 2022b). Soil samples must be collected enormously to satisfy the sampling design for this detailed approach, especially for large-scale OPP. In addition to these complications, laboratory-based N determination is laborious and time-demanding. Moreover, the process requires various chemical extractants, some of which are hazardous to the environment.

Predicting soil N in OPP using a regression model based on pedotransfer functions/PtFs could be a solution to a relatively quick N estimate at a specific location without relying too much on tedious and arduous work, as well as the potential for environmental degradation. PtFs approach is classified as data-driven science (Wadoux 2020a) and can be defined as functions developed from easily, routinely, or cheaply observed/determined soil properties to estimate specific soil properties that are infeasible to obtain (Bouma 1987; Padarian et al. 2018; Wadoux et al. 2021). The modern PtF approach performs statistical learning as its main regressor algorithm. Statistical learning refers to a broad set of methods that can be used to understand the data. One widely known example is machine learning/ML. ML algorithms can learn the diverse patterns of the data and use the stored previous information to improve their performance when learning and predicting new data (Jordan & Mitchell 2015; James et al. 2021). Furthermore, the ML algorithm is unconstrained by the requirements of underlying theories or pre-specified assumptions (*e.g.,* normality distribution; Wadoux

et al. 2020b) when making the prediction. This flexibility makes ML-based PtFs capable of capturing linear and non-linear natures of soil variables' interrelationships (Rossiter 2018; Padarian et al. 2020a). Currently, ML-based PtFs are widely used in predicting soil properties, such as soil texture (Liu et al., 2020), soil temperature (Abimbola et al. 2021), aggregate stability (Boushilim et al. 2021), soil organic C/SOC (Sakhaee et al. 2022), cation exchange capacity (Ma et al. 2021), soil reaction (Baltensweiler et al. 2021), available P (Kaya et al. 2022) and other soil macro- and micronutrients (Shi et al. 2022; Sihag et al. 2019). Eventhough the research for cultivated tropical peatlands was scarce (*e.g.,* Pulunggono et al. 2022a), PtFs employed with ML algorithms have been successfully implemented in estimating N in a maize field (Mashaba-Munghemezulu et al. 2021), rangelands (Parsaie et al. 2021), paddy soils (Xu et al. 2021) and other various land uses (Forkuor et al. 2017; Hounkpatin et al. 2022; Liu et al. 2022) which is entirely located in mineral soils.

In a similar location to this study, Pulunggono et al. (2022a) initially tried regression tree/RT-based PtF models for OP-cultivated tropical peat total N, P, and K contents. However, their report only described model performances and interpretation of a single type of tree-based ML algorithm and its comparison to multiple linear regression/MLR. As stated by Trontelj ml. & Chambers (2021), the RT algorithm is easy to implement and captures non-linear relationships among covariates. Unfortunately, the RT-based PfF model's low predictive performance may arise from smooth, gradual changes in the continuous covariates, as can be observed in Pulunggono et al. (2022a) results. Furthermore, this present study hypothesized that using more advanced ML algorithms such as ensemble trees (*e.g.,* random forest/RF, gradient boosting machine/GBM, and extreme gradient boosting/XGB), which are considered more complex than RT algorithm, thereby can handle previous constraints and may improve N predictive performance. RF, GBM, and XGB outperformed RT algorithm in mapping coastal wetlands (Wen and Hughes 2020) and predicting SOC (Goydaragh et al. 2021), Al and base saturation (Pulunggono et al. 2022c), as well as heterotrophic CO_2 emission from tropical peat (Pulunggono et al. 2022d) and dead root (Pulunggono et al. 2022e) with considerable accuracy.

One of the significant drawbacks of employing advanced and complex ML algorithms in PtFs-based modeling is their lack of interpretability and explainability, which are considered "*black boxes*". The lack of insight concerning model functioning and structure often emerges, which raises difficulties in discerning model information into the established pedological knowledge or soil biogeochemical processes (Rossiter 2018; Wadoux et al. 2020a; Wadoux et al. 2020b); *i.e.,* average and individual observation-based model predictions (Wadoux et al. 2020b; Wadoux and Molnar 2022), and the direction of dependent variable in responding the covariate's relative effects (Jones et al. 2022; Wadoux 2023). Often, the interpretability approach is developed explicitly to assess a particular algorithm, such as mean decrease in impurity and Gini index for tree-constructed models and its derivatives, *e.g.,* RT and RF; Breiman et al. 2002) and neuron connection weighting via Garson's or Olden's algorithms for artificial neural networks/multilayer perceptrons (ANN/MLP; Garson 1991; Goh 1995; Olden and Jackson 2002). Unfortunately, both approaches narrow the interpretation to a single or group of several algorithm families, thereby restricting a unified pedological interpretation and explanation of the entire diverse models.

One of the approaches developed to tackle the black box problem of ML algorithm is interpretable machine learning with a global approach regardless of the algorithm's types or model agnostic interpretation (Ribeiro et al. 2016). However, a local approach that can explain the model response at a given point of the predictor cannot be ignored since the global coverage often fails to satisfy a particular mechanism that often occurs in soils. One of the desirable tools for interpreting ML models enabling multi-approaches is SHapley Additive exPlanations (SHAP; Lundberg and Lee 2017), an approach to quantify a Shapley value in Game Theory (Shapley 1953). SHAP is widely known for its strong alignment with human intuition (Lundberg and Lee 2017). SHAP has been employed in explaining ML models in several soil studies, *e.g.,* soil organic C mapping (Padarian et al. 2020b), liquefaction potentials (Jas and Dodagoudar 2023; Sui et al. 2023), soil moisture (Huang et al. 2023), available P (Hall et al. 2023) and permeability coefficients (Tran 2022).

Therefore, this paper tried to (1) fit six pedotransfer models that consisted of general linear models/GLM and machine learning/ML to estimate total N distribution in peat drained and cultivated for oil palm/OP, (2) compare the ML calibration and validation performances against GLM-based pedotransfer models, and (3) explain and interpret the model functioning, primarily, key drivers in predicting total N distribution.

MATERIALS AND METHODS

Source of The Data

This study utilized the tropical peat nitrogen (N) dataset from Pulunggono et al. (2022a). The study area is oil palm/OP cultivated peatlands in Buatan Village, Koto Gasip District, Siak Regency, Riau Province, Indonesia, as shown in Figure 1. The soil dataset consisted of 120 data with an unbalanced design, taken at two sites, in the OP plantation (90 data) and its surrounding bush area (30 data). The bush is considered to be affected by OP management, particularly in hydrology. Three factors were presented, which comprised land use (OP and OP+bush), depth of sample collected (0-20, 20-40, and 40-60 cm), and distance from the oil palm tree (1.5, 3, and 4.5 m). However, this study only used the depth of the sample collected as a covariate due to the lack of data coverage of the other two factors. Furthermore, we selected three chemical properties (pH, organic C, and ash content) and four physical properties (particle density/PD and bulk density/BD) as the soil covariates, which had a total N.

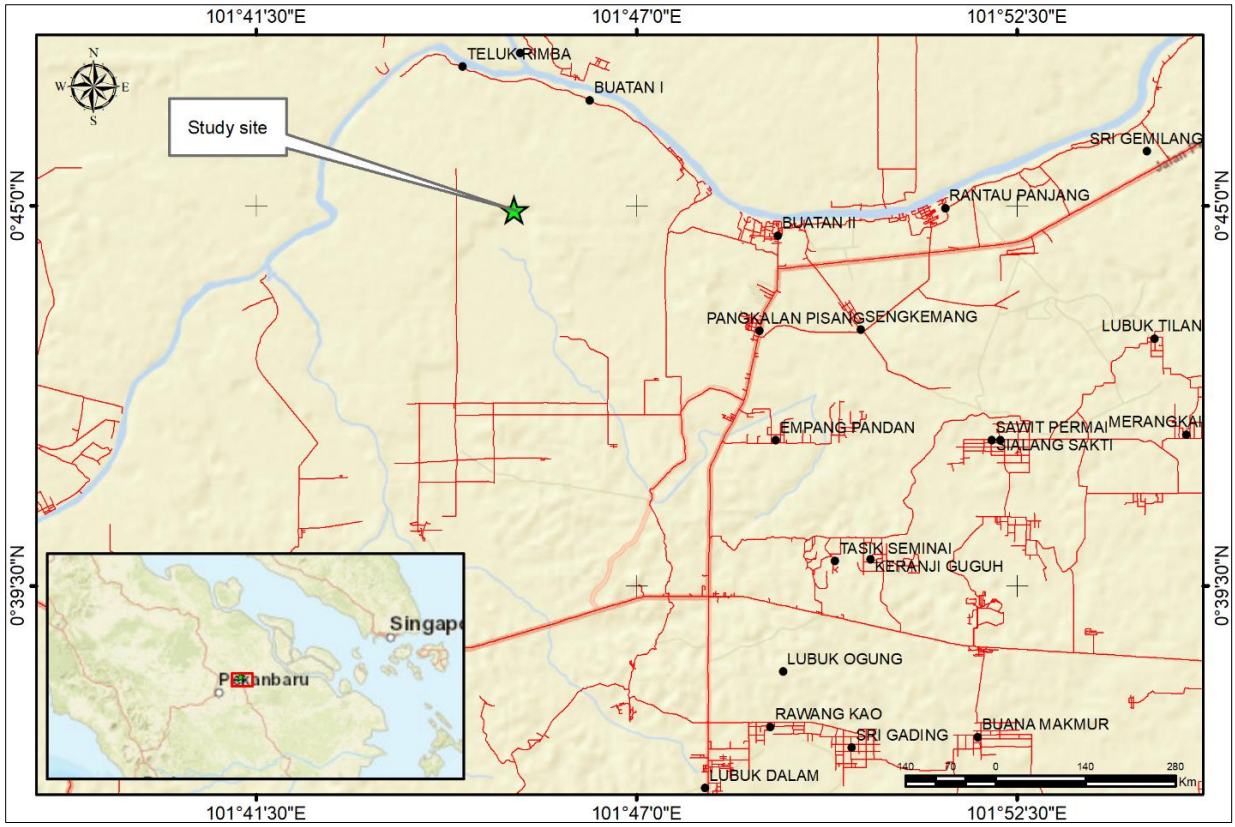

Figure 1. Location of the studied site

Model Development

As stated before, our dataset was referred from Pulunggono et al. (2022). Before being fed into the models, the dataset (N=120) was checked manually, which omits its missing values. Then, the dataset was bootstrapped (N=500) with replacement, resulting in 609 subsamples. This larger dataset was rechecked to match its distribution to the original dataset. Moreover, the bootstrapped datasets were then randomized (`seed 42`) and split into training and validation data. The trained data utilized 70% of the dataset (N=428), leaving the other 30% (N=181) as validation data. All the information regarding the original dataset is presented in Appendix 1.

Furthermore, the models were developed mechanistically involving environmental factors and soil properties that following the formula in Equation 1:

$$Total\ N \sim LU_{OP} + OP_{Dist} + Depth + Org\ C + Ash\_Content + CN\ ratio + pH + Total\ Fe + PD + BD + Porosity$$

.. Eq. 1

where: LU_OP, OP_Dist, Depth, and Org C represented land use, distance from oil palm tree, depth of sample collection, and organic C, respectively. One hot encoding technique was applied to LU_OP and OP_Dist since they present categorical values. LU_OP consisted of 0 and 1, representing bush and OP, respectively. The covariate selection process involves correlation analysis with a cut-off value of |0.7| and scatter plots to eliminate multicollinearity and identify the possibility of non-linear relationships. Both analyses were performed using `ggpair()` function from `ggally` package (Schloerke et al. 2023).

This study employed four rule- and ensemble-based tree learners commonly used in other studies: Cubist, random forest/RF, gradient boosting machine/GBM, and extreme gradient boosting/XGB. Modifying the M5 Model Tree idea from Quinlan (1992; 1993), Cubist first extracted a simplified rule set from a single decision tree. To strengthen its predictive power, Cubist employs a sequence of rule-based trees and post hoc prediction adjustment from the nearest neighbor in the original training dataset (called "committees" and "neighbor", respectively). With this regard, each tree is developed from the previous information to compensate for the prediction error. RF algorithm performs bootstrap aggregating and random subspace methods, which resample the training data and select particular covariates in random terms to build multiple fully grown trees. An individual deep tree has a higher variance by overfitting its training data. The RF algorithm minimizes this issue by averaging all the constructed trees to obtain model prediction (Ho, 1998; Breiman, 2001). GBM and XGB, in another way, perform boosting techniques, which sequentially construct shallow trees (also named "stumps"). A single stump is built similarly to RF using random resampling; meanwhile, it is considered a weak learner due to its over-generalization. Differently from RF, every successive stump was grown like committees in Cubist. Both algorithms minimize errors at each turn using stochastic gradient descent, which optimizes the model performances (Friedman, 2001; Friedman, 2002). Furthermore, XGB optimized the tree-boosting algorithm by introducing a particular approach to handle sparse data and using parallel and distributed computing (exploiting the out-of-core computation). This results in learning quickly from a large-scale dataset with accurate predictions (Chen and Guestrin, 2016).

Moreover, all the ML models were compared to the classical generalized linear models/GLMs and their derivatives, such as multiple linear regression/MLR, logistic regression (Log-GLM), and multivariate adaptive regression spline/MARS (as a more sophisticated polynomial-based model). All GLMs-based models were trained with a `caret` package (Kuhn 2023) using five-fold cross-validation/CV, repeated ten times. Furthermore, ML models were trained iteratively and exchangeably using `caret` and algorithm-specific packages, *i.e.,* `earth` (Milborrow 2023), `ranger` (Wright et al. 2023), `gbm` (Greenwell et al. 2022), and `xgboost` (Chen et al. 2023) with similar CV configuration. Optimum hyperparameter settings for MARS, RF, GBM, and XGB models are shown in Table 1. Meanwhile, the table did not include MLR and Log-GLM models since no particular configurations were applied to their algorithms.

Table 1. Model parameterization for all final models used in this study

Algorithm:	MARS	Cubist	RF	GBM	XGB
Package:	caret; earth	caret; Cubist	randomForest; caret	caret; gbm	caret; xgboost
Hyperparameter Settings:	degree: 3 nprune: 23	commitees:100 neighbors:1	mtry: 11 ntree: 500	n.trees: 6000 interaction.depth: 3 shrinkage: 0.3 n.minobsinnode: 3	nrounds: 10000 max_depth: 4 colsample_bytree: 0.8 eta: 0.1 gamma: 0 min_child_weight: 2 subsample: 0.5

Model Agreement

This study employed RMSE, MAE, and R^2 (coefficient of determination/Rsquared) to evaluate the model performance by inspecting the agreement between the predicted and observed total N values. This study adopted two methods, *i.e.,* performance-based (1) calibration and (2) validation data. The first method (executed using `caret::resamples()` function) took advantage of available performance data from the repeated CV fold (N=50) that are computed internally during the training process. The last method was executed manually using 30% validation data. We added a bias metric to point out the average direction of model prediction towards their observed values using `modelmetrics` package (Hunt 2022). The entire specific information for all model agreements is available in Appendixes 2 and 3.

Model Interpretation: Shapley Additive Explanation (SHAP) Value

Game Theory generally studies the mathematical explanation of games and their players' or agents' collaborative interaction and strategies (Nash 1950; Rasmusen 1989). Within the context of modeling, game theory posits that the predictive model is a confluence of multiple collaborative games. In each game or prediction attempt, players (or covariates of the model) strategically collaborate to achieve a specific goal, which in this context refers to a predicted value. In exchange for their collaborative efforts, each player receives a "share" or "payout" corresponding to their individual contribution. The deviation from the mean prediction for the complete dataset determines the collaborative total loss or gain.

The Shapley Additive Explanation, or SHAP, was first described in Lundberg and Lee (2017) paper and provides a conditional expectation function to calculate the Shapley value (Shapley 1953). The Shapley value quantifies an individual player's contribution to a coalition, representing the expected marginal contribution of the covariates across all possible combinations. Following Padarian et al. (2020) and Molnar (2023) explanations, the Shapley value can be computed by fitting the model $f_{S\cup\{i\}}$ involving the covariate i, while withholding covariate i in another model, denoted as f_S. Covariate i's marginal contributions is calculated using the difference between the model prediction on the input x (or mathematically expressed as $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$). This process is repeated iteratively, considering the complete power set of the covariates (including all possible subsets of $S \subseteq F$ and the set F) to obtain the weighted average of all marginal contributions ($\phi_i \in \mathbb{R}$), as summarized in Equation 2.

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right]$$

.. Eq. 2

However, due to its prohibitive nature in handling large amounts of covariates or using thousands of simulations, for the more complicated model, we must re-train computationally exhaustive $2F^{|F|}$ models, the Shapley value can be approximated by SHAP. This method simplifies the Shapley value estimation using sampling approximation derived from the model's training dataset, accelerating the computation. According to Lundberg and Lee (2017), SHAP is based on using g as a simpler explanation model in explaining f as original prediction model. Following the concept of local explanation, a prediction model $f(x)$ can be explained using simplified input x' that map an original inputs using a mapping function $x = h_x(x')$. This local approach ensures $g(z') \approx f(h_x(x'))$ while it applied by different set of inputs $z' = x'$ (Equation 3).

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i'$$

.. Eq. 3

The inputs z' are derived from the training dataset using a particular resampling procedure. Furthermore, the marginal effect ϕ_i to each covariate (Equation 2) is supplied to Equation 3 as a linear function of binary variables ($z' \in \{0,1\}^M$), where M is the number of simplified input features and the

number of $\{0,1\}$ represents the availability (available or not) of covariate contribution. The prediction of model $f(x)$ can be approximated by summing the effect of all covariate contributions in Equation 3. Furthermore, SHAP provides a missingness solution for Equation 1, where $f_x(z') = f\big(h_x(z')\big) = E[\,f(z) \mid z_s\,]$, and S is the set are non-zero indexes in z', whereas $f(z_S) = E[f(z) \mid z_s]$ for handling missing values (z_s has missing values for covariates not in the set S), as their ordering calculations according to Lundberg and Lee (2017), were depicted in Figure 2.

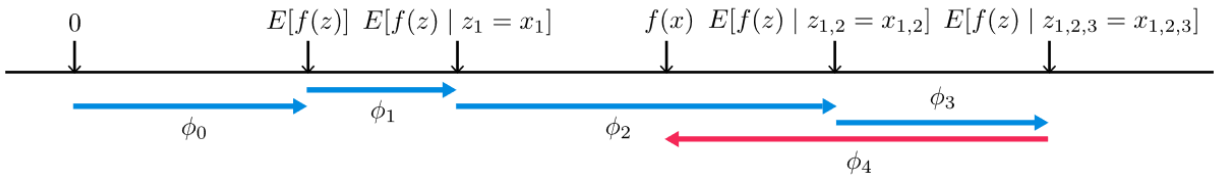

Figure 2. The graphical concept of mathematically quantifying and ordering SHAP values for each covariate when conditioning on that covariate. This figure is derived from Lundberg and Lee (2017) paper.

Quantifying SHAP for local and global explanation via Shapley value was attained using `fastshap` and `shapviz` packages (Greenwell 2023; Mayer and Stando 2023) in the R environment. The trained models were first explained using the `fastshap::explain()` function. To provide accuracy and stability of the explanation, 1000 Monte Carlo resampling technique (`nsim = 1000`) iteratively computed to supply the equation 2, following trumbelj & Kononenko (2013) approach (Greenwell 2023; Molnar 2023). Then, the explaining information was transferred to the `shapviz` object, excluding the MARS-based PtFs model. Since the fundamental of MARS is the automatic covariate selection to fit the dataset's underlying distribution, forcing all covariate predictions using the Shapley value is impossible.

The global explanation was computed using the training dataset and visualized using `sv_importance()` function, presented by covariate importance and beeswarm plots. The first plot employs the mean Shapley value by averaging all individual contributions; meanwhile, the latter depicts all individual contributions in a single plot. The local explanation was evaluated using the prediction at `row_id=c(3,100)`, then visualized as waterfall plots using `sv_waterfall()` function. Furthermore, the (expected) covariate marginal contribution was determined by utilizing the `sv_dependence()` function. This involved plotting individual Shapley values against the four most significant covariates of the PtF models, referred to as dependence plots. Notably, the strongest interacting covariates were represented through color gradients. To maintain cohesion and conciseness within the paper, only the most proficient GLM and ML algorithms were selected for visualization, providing a comparative analysis of their performance in predicting the total N in peat. However, the performances of other PtFs' models, as well as the associated Shapley-based explanations, were extensively documented in appendices 5, 6, 7, and 8 for supplementary reference.

RESULTS

Dataset Description and Covariate Relationships

The results of the peat chemical analyses were summarized in Figure 3, distinguishing between the training and validation datasets. The figures revealed that the chemical variables in both datasets fell within the typical range observed in tropical peat, as reported by previous researchers (*e.g.,* Adeolu et al. 2018; Dhandapani et al. 2021; Harianti et al. 2018; Hashim et al. 2019; Pulunggono et al. 2022b; Sangok et al. 2017). However, the recorded total Fe in this study was three times lower than that observed by Dhandapani et al. (2020) in similar land use. The frequency distribution of total N in the training dataset mirrored that of the validation dataset, as indicated by the presence of bimodal peaks in Figure 3. Additionally, other soil covariates displayed similar patterns.

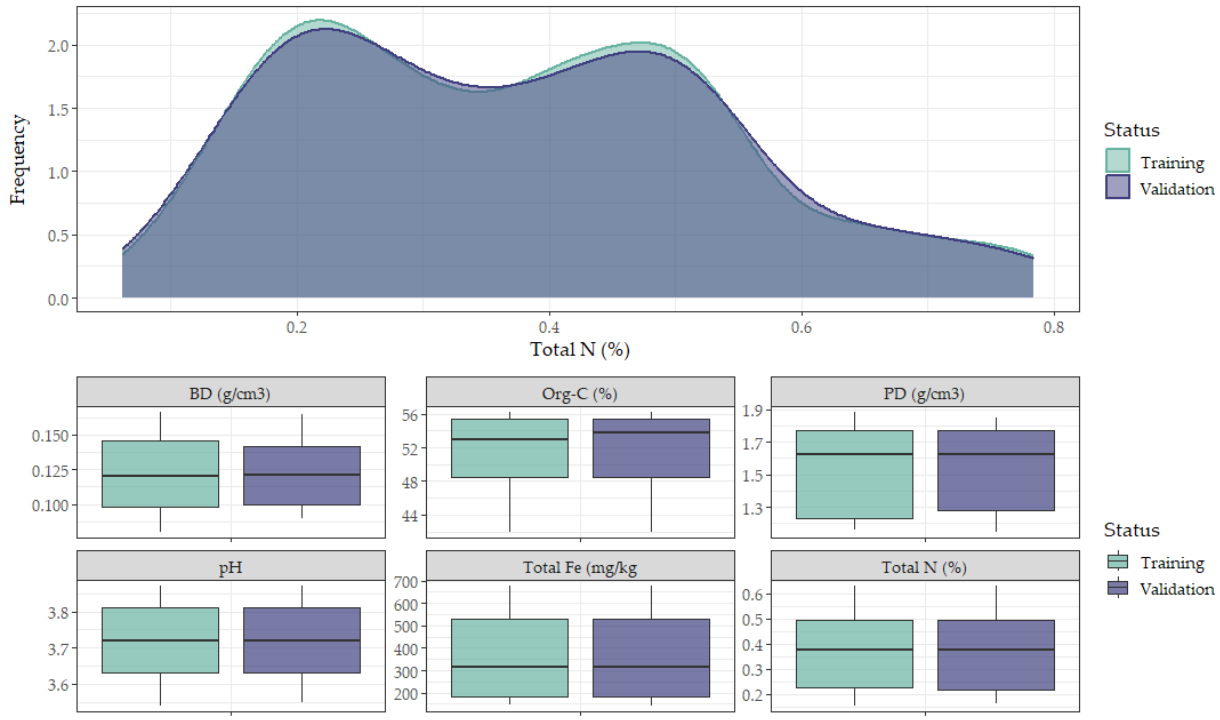

Figure 3. All covariate distributions are divided by training (green, hex:#69b3a2) and validation (purple, hex:#404080) dataset. All presented soil covariates were the final result after the collinear covariates were eliminated.

Figure 4 depicted the relationships between variables, showcasing that total N exhibited a moderately significant positive linear relationship with sampling depth, while its relationship with total Fe was low and insignificant. Conversely, total N demonstrated a significant negative correlation with organic C. Similar but insignificant trends were observed for pH, bulk density (BD), and particle density (PD).High correlations were found between organic C against CN ratio and ash content and BD and porosity at the initial model training (>|0.7|; data not presented), which may introduce high multicollinearity. Considering their weak relationship with total N, CN ratio, ash content, and porosity were dropped from the PtF models. All low to moderate correlations presented in Figure 4 showed the possibility of non-linear relationships between total N against environmental factors and soil covariates.

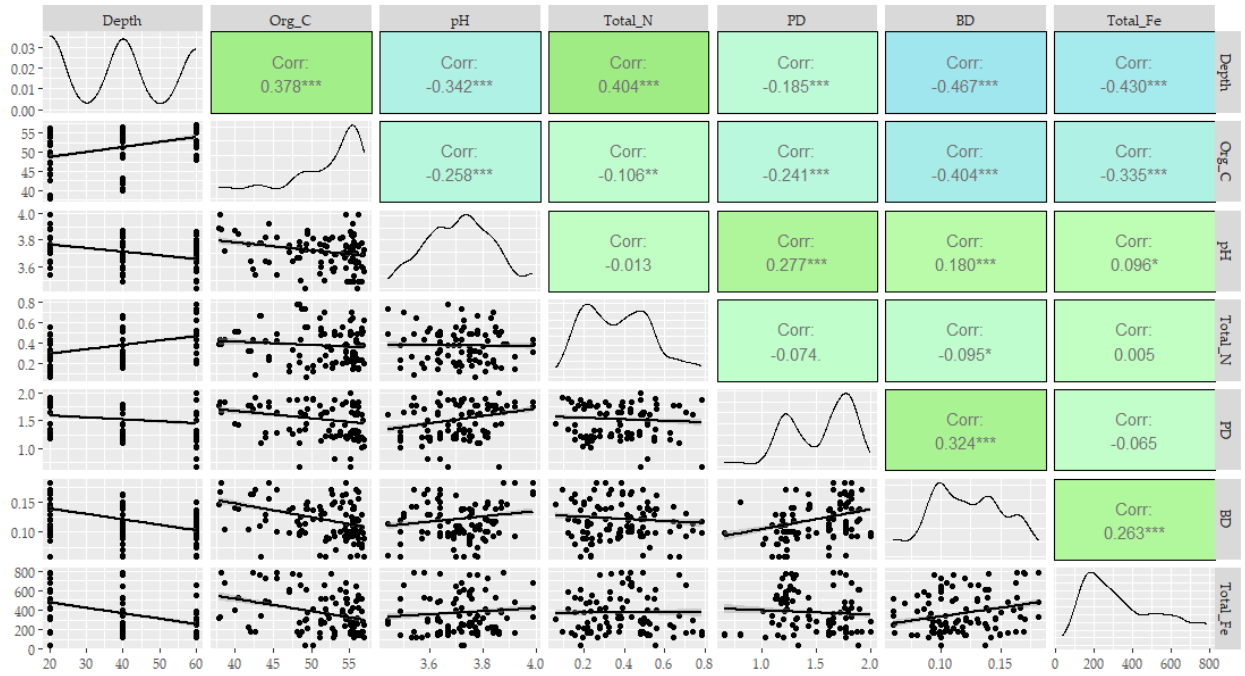

Figure 4. The relationships between each covariate used in this study

Pedotransfer Model Agreement

This study found that the Rsquared, RMSE, and MAE of all PtF models were diversely varied. All ML algorithms outperformed GLM-based PtF models. It could be seen on a relatively higher Rsquared (> 0.90) and lower RMSE (<0.04) and MAE (<0.02) of ML-based PtF models computed from the calibration method (Figure 5). Validation agreements also showed similar metric patterns with higher performances (Table 2).

As shown in Figure 5 and Appendix 3, GBM and XGB algorithms achieved remarkable and stable agreements near the perfect scores (*e.g.,* Rsquared > 0.98). Both regressors also possessed the lowest RMSE and MAE compared to other algorithms, whereas XGB outperformed GBM regarding median and mean values in both metrics. The Cubist algorithm also yielded high-performance metrics comparable to GBM and XGB and better than RF. However, it had unstable performances, as can be observed by higher variability throughout the greater interquartile length compared to GBM and XGB. RF had good agreement (Rsquared = 0.97) and lagged behind GBM, XGB, and Cubist. Consequently, it achieved the worst performance compared to other ML algorithms but was still more stable than Cubist and considerably higher than all GLMs. MARS was the best GLM regressor. It attained a moderate agreement, which accounted for half of ML accuracies. Contrastingly, LM and LogGLM were the worst regressors evaluated by all metrics.

All the ML algorithms had more comparable predictions when tested using validation data (Table 2), especially for the Rsquared metric. Cubis outperformed all other ML-based PtF regressors concerning all the evaluated metrics. GBM and XGB had comparable metrics, similar to Cubist in Rsquared but remarkably lower in three other metrics. Similar to the calibration method, GLM-based PtF models yielded the lowest Rsquared. Particularly for the MARS algorithm, it scored the worst Rsquared (R^2=0.17) and comparable RMSE and MAE to other GLM-based PtF regressors (both 0.15 and 0.13, respectively). As negative bias shows, RF and GBM underestimate the total N values over their actual values. Oppositely, the other algorithms attained contrasting results.

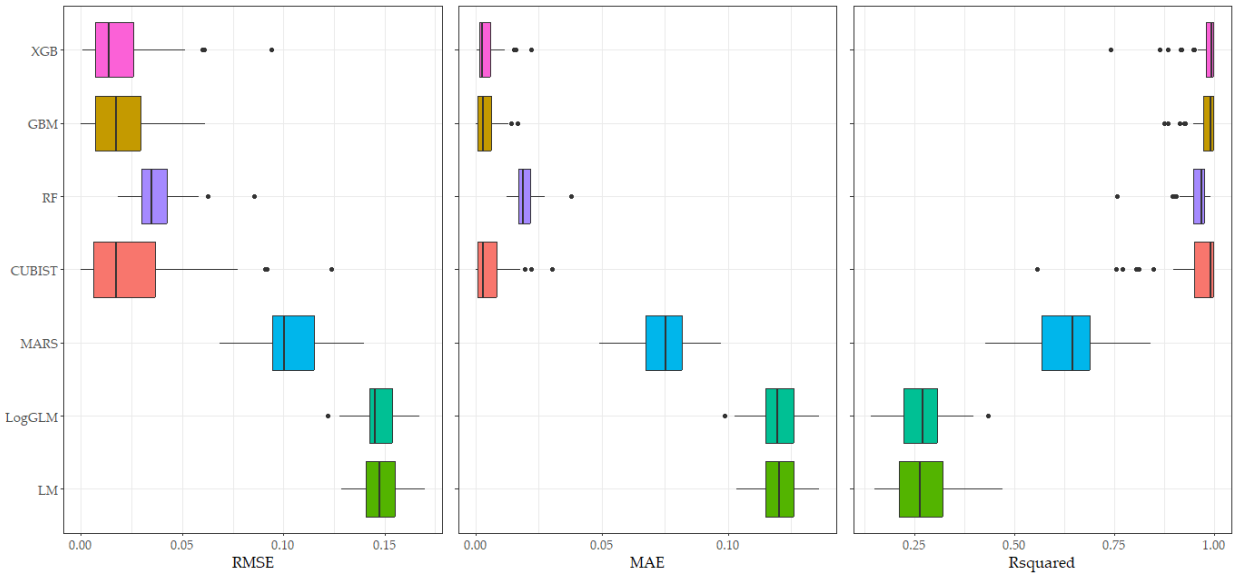

Figure 5. Calibration agreement of six GLM and ML-based PtF models, computed from repeated five-fold CV during the training process (N=50)

Table 2. Validation agreement of six GLM and ML-based PtFs models, computed using 30% of the dataset

Model	Rsquared	RMSE	MAE	BIAS
MLR	0.329	0.140	0.115	9.12×10^{-5}
LogGLM	0.337	0.139	0.114	-1.06×10^{-4}
MARS	0.720	0.091	0.064	-4.94×10^{-3}
CUBIST	0.999	1.21×10^{-7}	2.33×10^{-8}	6.71×10^{-9}
RF	0.992	1.68×10^{-2}	7.87×10^{-3}	1.40×10^{-3}
GBM	0.999	1.59×10^{-4}	1.02×10^{-4}	5.98×10^{-6}
XGB	0.999	1.83×10^{-3}	8.72×10^{-4}	1.54×10^{-4}

Explanations of Pedotransfer Models based on SHAP

Figure 6 shows the PtF models' global explanations by visualizing differences in taking the important covariates between LogGLM, GBM, and Cubist algorithms based on Shapley values, depicted in their averaged and individual covariates contributions. Except for RF, all models agreed on sampling depth, organic C, and total Fe as their most important covariates, regardless of contrasting prediction accuracies (Figure 6, Appendix 5). LogGLM seemingly relied heavily on depth as its first important covariate, while other covariates had low to negligible strengths. This opposes GBM and Cubist, which also reserved total Fe, BD, and pH as additional substantial covariates with comparable moderate contributions. Unexpectedly, we detected weak importance of land use and distance from the OP tree on the best performer PtF algorithms (RF, Cubist, GBM, and XGB), while the latter covariate showed up at GLMs (Figure 6, Appendix 4).

The individual contribution in beeswarm plots in Figure 6 depicted that in particular ranges of their values, all important covariates in both algorithms might negatively contribute to the total N prediction while also resulting in positive contributions in different ranges. For instance, in the GBM algorithm, a negative to low contribution is obtained when the peat is sampled over 40 cm and the peat materials possess a relatively higher organic C. Oppositely, deeper sampling, and lower organic C content primarily contribute to positive Shapley values. This pattern was seemingly amplified in the Cubist model, particularly for higher depth. Other covariates also exhibited a similar pattern, especially in total Fe. This covariate substantially contributed to the negative and positive Shapley

values at lower and higher concentrations, respectively. However, a low contribution is obtained at total Fe's moderate content (Figures 6, 7, and 8). An interesting pattern was also observed in the land use covariate as a less critical covariate, highlighting a substantial negative contribution in the absence of OP (Figure 6).

The local explanation shown in Figure 7 decomposed the PtF model's strategy in estimating total N using covariate information available in `row_id = c(3,100)`. It can be observed that the behavior of PtF algorithms deviates from their general explanations, as captured in Figure 6. In GBM, the important covariates were rearranged, with total Fe and sampling depth mainly and diametrically contributing to the total N prediction. Nevertheless, all notable contributions were supplied by important covariates as generally presented in Figure 6 (*i.e.,* total Fe, sampling depth, pH, and BD) except for organic C. As shown in Figures 6 and 8, the likelihood of low contributions of organic C to the N prediction could be detected at higher values (more than 54%).

The dependency plot in Figure 8 presented an individual marginal contribution of the four most important covariates (*i.e.,* sampling depth, organic C, total Fe, and BD) on the total N prediction, along with their strongest interacted covariates. LogGLM fitted predictable linear relationships on both covariates; meanwhile, GBM captured non-linear patterns. It could be observed that the Shapley values responded with a flat positive slope towards a sampling depth over 40 cm. The relationship had a steep positive slope at deeper sampling in OPP, indicating a Shapley value's strong response. Organic C generally had a contrasting slope direction against sampling depth, measured using LogGLM and GBM algorithms. In LogGLM, low organic C and high PD contribute positively to total N prediction, whereas opposite patterns exhibited in contrasted both covariate's values. The marginal contribution relationships of GBM for organic C, total Fe, and BD were more complicated and thus suggested fluctuating patterns; however, they follow similar general trends with LogGLM.

We observed that the peat's seemingly physical properties were the invariable counterpart of organic C, judged by its appearance as the strongest interacted covariates on both contrasting algorithms. Other algorithms also exhibited similar relationships, including LM and RF (Figure 8; Appendix 7).

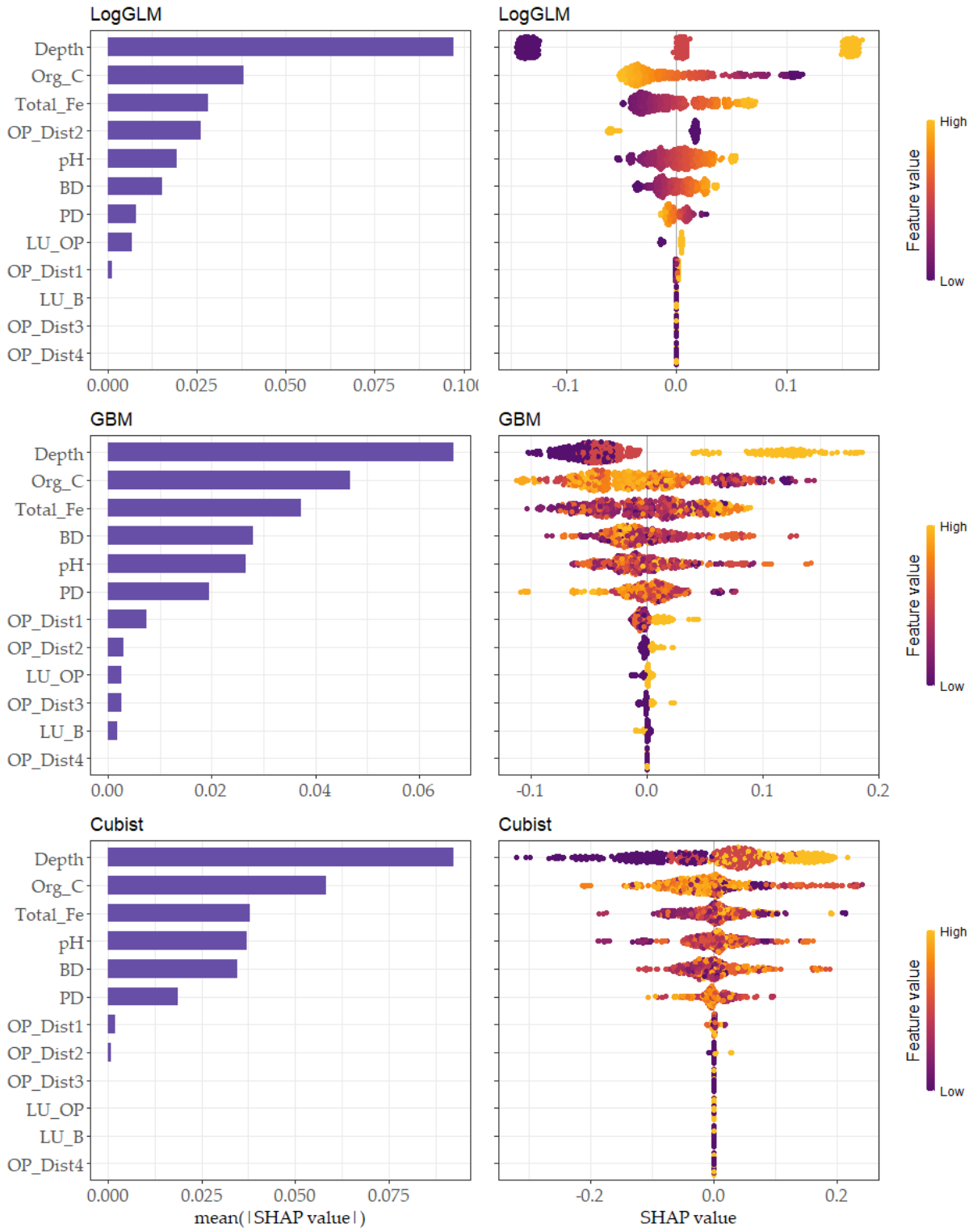

Figure 6. SHAP-based Global explanation covariate contributions that were captured by averaging all contributions of each covariate (left) and all individual Shapley values (right), calculated according to LogGLM (above), GBM (middle), and Cubist (bottom), as the best GLM and ML-based PtF regressors, respectively

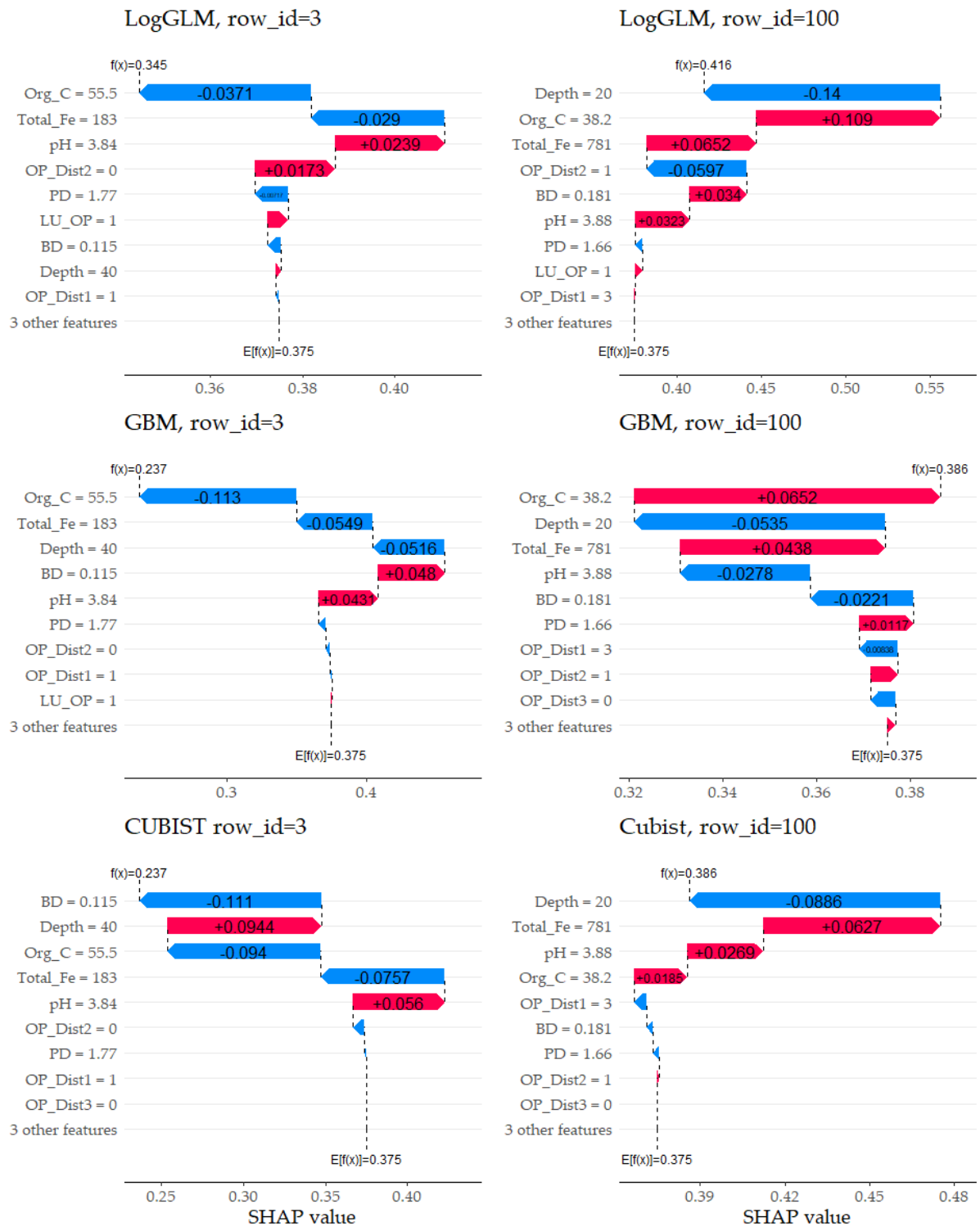

Figure 7. SHAP-based Local explanation captured by covariate contributions to an individual prediction on `row_id = 3` (left) and `row_id = 100` (right) based on Shapley values, calculated according to LogGLM (above), GBM (middle), and Cubist (bottom) as the best GLM and ML-based PtF regressors, respectively. Note that $f(x)$ denotes the prediction on the SHAP scale, while $E(f(x))$ refers to the baseline SHAP value.

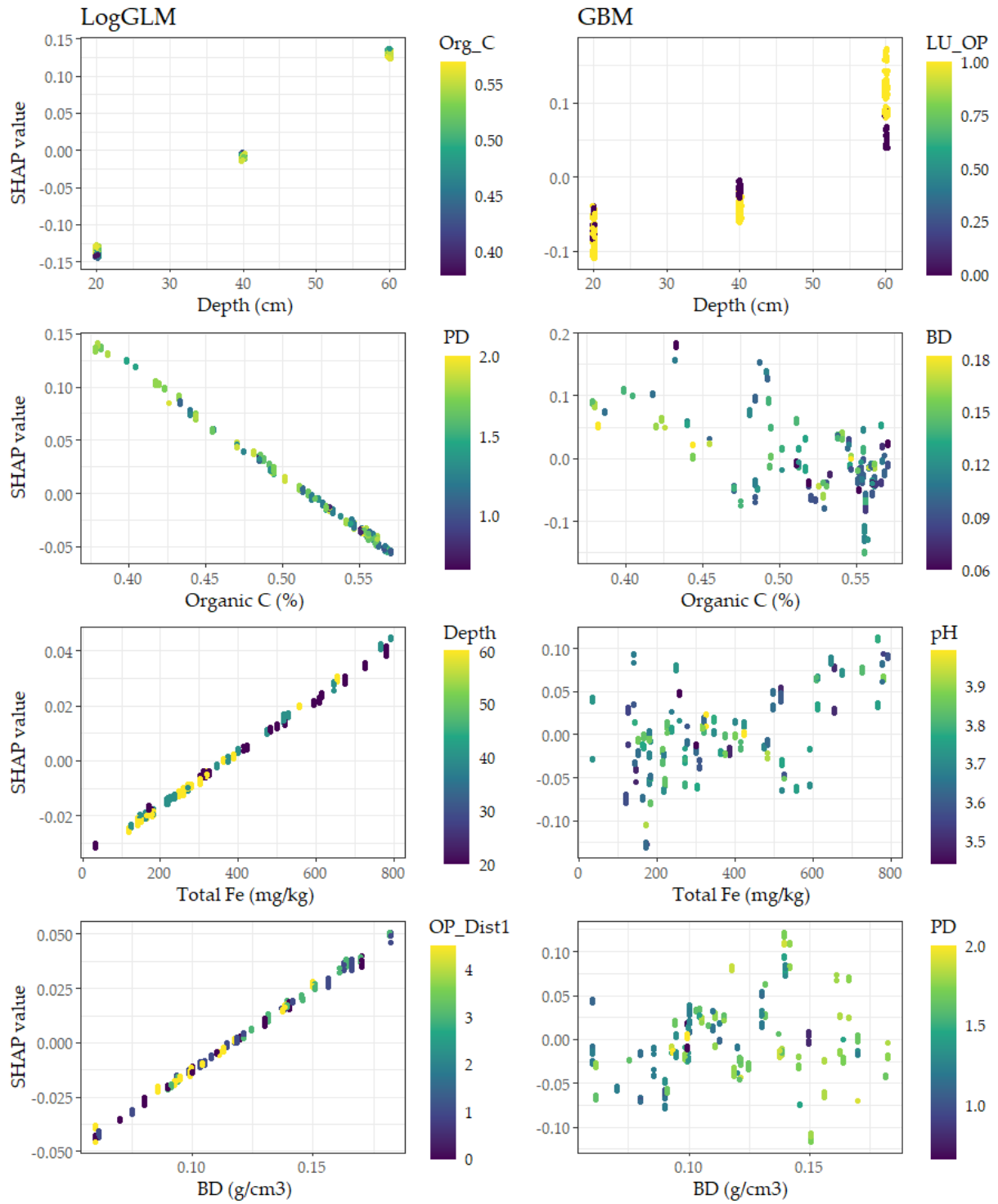

Figure 8. The dependency plot of SHAPs against the four most important covariates is calculated according to LogGLM (left) and GBM (right) as the best GLM and ML-based PtF regressors, respectively. The strongest interacting covariates supplied the color gradients.

DISCUSSION

Intercomparison between ML-based PtF Model: Insight from Robust Statistical Learning

This study demonstrated that using datasets from Pulunggono (2022a), the performances of ML-based PtF models were satisfactory in estimating total N in OP-cultivated tropical peat (Figure 5 and Table 2). Based on the calibration (Figure 5) and validation methods (Table 2), the best performer algorithms for PtF models were GBM, XGB, and Cubist. The agreement metrics of best-performer algorithms were relatively higher than those reported by Song et al. 2020 in predicting the total N in mineral soils. Furthermore, our GLM-based PtF modes exhibited low agreements, contrasting with Mesele and Ajiboye (2020) and Karim (2023) studies. This small-scale study highlights the potential to use artificial intelligence to predict nutrients in tropical peat drained and cultivated by OPP, particularly at the spatial and temporal scales.

Our results partially showed that simple Ptf models (*i.e.,* based on GLM) could suffer from several possible confounding factors, *i.e.,* complex and non-linear relationships, covariate gaps, and superficial/unflexible algorithmic design. The constraint introduced by the combination of observation is commonly found in predictive modeling worldwide. Especially in predicting soil properties, higher sample sizes had a vital role in maximizing the model's representation of natural variance (Maharana et al. 2022; Xu and Goodacre 2018). Most PtF backed with ML algorithms are trained using large samples (*e.g.,* Baltensweiler et al. 2021; Liu et al. 2020; Moghaddam et al. 2020; Shi et al. 2022). Due to financial and accessibility constraints in many peatland studies, the sample size is often inadequate to develop a robust predictive model. Low agreement of ML-based PtF models due to low observation constrained by unflexible algorithmic design can be observed in the Pulunggono et al. (2022a) study.

Furthermore, Rajput et al. (2023) and Lindstromberg (2023) reviewed that undersampling can lower the probability of producing true effects, enhancing false-negative results, and exaggerating the winner's curse effect. We overcome this limitation by applying a bootstrapped resampling technique, which generates a particular Monte-Carlo simulation for random subsampling with a relatively similar underlying distribution to the original dataset (Figure 2; Appendix 1). Nevertheless, we acknowledge that our resampling and data-splitting techniques might be biased, and the propagated errors may result in lower accuracy in the field deployment compared to their performance in the current report. Both techniques assume independent observations, which consequently precludes spatial autocorrelation (Efron 1979; Efron and Tibshirani 1994; Karasiak et al. 2021; Lyons et al. 2018; *e.g.,* Iranpanah et al. 2003; Tang et al. 2006) that unaccounted for our environmental factors.

Eventhough all bootstrapped samples, including their dependent and independent covariates, had similar distribution compared to the original dataset ((Figure 2; Appendix 1) and possessed significant covariates (Appendix 2), our GLM-based PtF models failed to accomplish comparable performances towards tree-based ML regressors (Figure 5; Table 2). The summary of the LM model indicated that nearly all covariates significantly influenced total nitrogen (N), except land use difference and particle density. Conversely, the Log-GLM produced unsatisfactory results in identifying influential factors, indicating a lack of logistic-based relationships between the total N value, the studied factors (encoded as binary variables), and soil covariates (Appendix 2). It can be observed that under a high sample size, the ML-based PtF algorithm yielded near-perfect scores. This condition was also observed in validation-based performance since the validation dataset resembles those used for training (Figure 2). Since the underlying mathematical nature of bootstrapping neglects outliers, it may quickly lessens the learning pace of the machine learners, but not for GLMs due to its rigid OLS estimator. These results may occur due to delicate interrelationships introduced by the combination of environmental factors and soil covariates (Figures 4, 6, and 8), which could be better captured by more complex and flexible statistical learning algorithms at large observation, *e.g.,* GBM or a computationally efficient algorithm such as XGB. Concern raised by several soil scientists (Rossiter 2018; Wadoux et al. 2020c) and researchers from other fields (McPherron et al. 2022; Lypton 2018), who look the noticeable and intriguing predictive ability that resulted from ML and its associated statistical methods (*e.g.,* bootstrapping, cross-validation, jacknifing, Monte-Carlo simulation and others) are must be taken with care by relating its pattern recognition to an established pedological knowledge.

In this study, we remark some insight for future research in PtFs-based nutrient modeling in peat as follows:

1) Supplying more flexible GLM-based algorithms for PtFs, for instance, partial least squares/PLS or principal component regression/PCR may remedy the GLM's performance problem, as suggested in other modeling studies (Li et al. 2020; Sarkar et al. 2023; Xu et al. 2018),
2) An appropriate approach for dataset resampling and splitting is necessary for insufficient sample size as well as on which the observation had spatial, temporal, or hierarchical dependencies (Karasiak et al. 2021; Lyons et al. 2018; Roberts et al. 2017). Despite its controversies (Young 1994; McPherron et al. 2022), other modeling studies embrace resampling as a tool to provide adequate input for data-hunger ML algorithms (Huang and Huang 2023).
3) Training RF, GBM, and XGB as the primary regressors using large sample size and covariates might require more sophisticated approaches in tuning their internal parameters (*e.g.,* applying a customized `tidy-modelling` instead of `caret`; Kuhn and Silge 2023) to maximize their full computational capability and flexibility. We consider Cubist the most convenient algorithm among other ensemble-based tree learners since it carries explainable information, a simple tuning process, and robust predictive power.
4) Substantial model performance as well as relational and pattern information based on ML algorithms, must be carefully handled, highlighting their interpretation from established pedological theory (Rossiter 2018) and position in knowledge discovery (Wadoux et al. 2020c).

Key Drivers Regulating Total N Prediction in PtFs: Bridging SHAP-based Findings with Pedological Explanation

Based on the SHAP method, our models, including the best performer ML-based PtF model, highlighted the higher importance of sampling depth and organic C in estimating total N in cultivated tropical peat (Figures 7 and 8). High-performance matrices also reported by Song et al. (2020) in predicting total N in China's mineral soils using LM-based PtFs and soil properties (SOC, CEC, and sand content), achieving Rsquared and RMSE that were ranging from 0.55 to 0.93 and 0.21 to 0.79, respectively. Training PtF models from DSM drastically decreased Rsquared to around 0.03 to 0.53. Using CEC, sand content, and topography for total N estimation in peat might be unuseful in our study area due to their homogeneous nature from a relatively close sampling site and similar peat thickness. In Ghanaian mineral soils, Mesele and Ajiboye (2020) developed eight successful PtFs (Rsquared 0.77 – 0.85; N = 120) in predicting total N using organic C as a covariate. Also, in mineral soils, Karim (2023) reported that polynomial- and linear-based PtFs could be estimated with higher precision (Rsquared = 0.82 and 0.81, RMSE = 0.18, respectively). Their results could not be applied to the OPP peat soils since we found contrasting results (Figure 5; Table 2).

Our results agreed with the previous findings that revealed the relationships of soil organic C to N supply in peat (Pijlman et al. 2020). Kononen et al. (2018) reported that microbial C and N were highest at the peat surface and decreased along with increasing depth. Based on their report, depth can be used as a proxy of N since high microbial occurrence indicates high decomposition potential. On the other hand, our SHAP study found a significantly contrasting pattern (Figures 4, 7, and 8), which may indicate higher N transportation and organic matter decomposition at the subsurface peat. Eventhough Hashim et al. (2019) suggested that shallow groundwater table generated higher N and other base cations subsurface leaching, the study site had mature OP and hydrologically managed, as well as periodic shallow water table during the rainy season (Pulunggono et al. 2023f). We argued that these conditions might allow the fertilizer-derived mobile N fractions to be transported downward near GWL (Rьckauf et al. 2004; Zhao 2001), and more homogeneous aeration throughout the upper 60-70 cm generates a higher decomposition at the lower profile.

Glendining et al. (2010) conducted a study that revealed the necessity for improving PtF predictors in accurately estimating total nitrogen (N) content in soils with more than 10% soil organic carbon (SOC). Their research, which had a larger sample size (N=265) than ours, also demonstrated that higher N content introduced significant variability. These findings align perfectly with the scope of our study. Notably, all peat samples analyzed in our study exhibited SOC levels exceeding 30% and higher N content, as depicted in Figure 3. Given that our study site is a relatively flat region at the river backswamps (Figure 1), topographic covariates may not contribute significantly to the PtF models, as they were not accounted for in our analysis. Additionally, our Shapley value analysis indicated that total N exhibits non-linear and complex interrelationships with the environmental factors and soil covariates, as illustrated in Pulunggono (2022a), as well as Figures 4, 6, and 8. These factors, coupled with the absence of substantial covariates exhibiting a linear

relationship with total N, resulted in lower performance of GLM-based models, as depicted in Figure 5 and Table 2.

This study accounted for the difference in land use in predicting total N. Song et al. (2020) found that the performance of PtF models decreased when trained from different landscape compositions. Differently from the established knowledge, at the local scale, our study neglected the contribution of land use difference (OP vs. bush) and distance from OP in predicting total N in drained OPP peat (Figures 6 and 7). The land use difference might induce these results since we use the dataset combining OP and bush data. This difference might occur due to varying degrees of decomposition stage and N supply from fertilizer. To maximize nutrient absorption at the highest root density, fertilization and amelioration are conducted on the outer boundary of the fertilization circle.

CONCLUSIONS

The predictive performances of ML-based PtF models in estimating total N content in cultivated tropical peat were found to be significantly superior to those of GLMs, using bootstrapped dataset from Pulunggono (2022a). The calibration and validation approaches revealed that the top-performing algorithms for PtF models were GBM, XGB, and Cubist, while the least effective models were LM and Log-GLM. The SHAP method analysis indicated that sampling depth and organic C were consistently identified as the most influential factors across all models, regardless of the predictive algorithm used. Additionally, ML algorithms highlighted total Fe, pH, and BD as additional significant covariates. Local explanations based on Shapley values suggested that the behavior of PtF-based algorithms might deviated from their global explanations. This study emphasized the crucial role of ML algorithms and game theory in accurately predicting total N content in peat that has been drained and cultivated for OPP, while also shedding light on the behavior of the models in relation to soil biogeochemical processes.

ACKNOWLEDGMENT

The Indonesian Oil Palm Plantation Fund Management Grant/BPDPKS supported this research. The authors thank Astra Agro Lestari Tbk's management and field staff for their kind support and assistance during the field campaign.

DATA AND CODE AVAILABILITY

The paper uses data obtained from Pulunggono et al. (2022a). All bootstrapped data could be available upon responsible request, whereas the original datasets analyzed during the current study are not publicly available. All modeling codes and updates associated with this current submission can be accessed on GitHub: https://github.com/fajrintanah/ML_TN_Peat.

REFERENCES

Abimbola OP, Meyer GE, Mittelstet AR, Rudnick DR, & Franz TE. 2021. Knowledge-guided machine learning for improving daily soil temperature prediction across the United States. *Vadose Zone Journal*, 20(5). https://doi.org/10.1002/vzj2.20151

Adeolu AR, Mohammad TA, Nik Daud NN, Sayok AK, Rory P & Stephanie E. 2018. *Soil Carbon and Nitrogen Dynamics in a Tropical Peatland. Soil Management and Climate Change: Effects on Organic Carbon, Nitrogen Dynamics, and Greenhouse Gas Emissions*, 73-83. https://doi.org/10.1016/B978-0-12-812128-3.00006-9

Anda M, Ritung S, Suryani E, Hikmat M, Yatno E, Mulyani A, & Subandiono RE. 2021. Revisiting tropical peatlands in Indonesia: Semi-detailed mapping, extent and depth distribution assessment. *Geoderma*, 402:115235. https://doi.org/10.1016/j.geoderma.2021.115235

Baltensweiler A, Walthert L, Hanewinkel M, Zimmermann S, & Nussbaum M. 2021. Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland. *Geoderma Regional*, 27:e00437. https://doi.org/10.1016/j.geodrs.2021.e00437

Bou Dib J, Krishna, VV, Alamsyah Z & Qaim M. 2018. Land-use change and livelihoods of non-farm households: The role of income from employment in oil palm and rubber in rural Indonesia. *Land Use Policy*. 76:828–838. https://doi.org/10.1016/j.landusepol.2018.03.020

Bouma J. 1989. Using soil survey data for quantitative land evaluation. *Advances in Soil Science*, 177–213. https://doi.org/10.1007/978-1-4612-3532-3_4

Bouslihim Y, Rochdi A, & El Amrani Paaza N. 2021. Machine learning approaches for the prediction of soil aggregate stability. *Heliyon*, 7(3):e06480. https://doi.org/10.1016/j.heliyon.2021.e06480

Breiman L. 2001. Random forest. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman L. 2002. *Manual on Setting Up, Using, and Understanding Random Forests v3. 1.*, California (US): Statistics Department University of California Berkeley. p.1:58

Broeshart H, Ferwerda JD, & Kovachich WG. 1957. Mineral deficiency symptoms of the oil palm. *Plant and Soil*, 8(4):289–300. https://doi.org/10.1007/bf01666319

Chaddy A, Melling L, Ishikura K, & Hatano R. 2019. Soil N_2O emissions under different N rates in an oil palm plantation on tropical peatland. *Agriculture*, 9(10):213. https://doi.org/10.3390/agriculture9100213

Chaddy A, Melling L, Ishikura K, Goh KJ, Toma Y, & Hatano R. 2021. Effects of long-term nitrogen fertilization and ground water level changes on soil CO_2 fluxes from oil palm plantation on tropical peatland. *Atmosphere*, 12(10):1340. https://doi.org/10.3390/atmos12101340

Chen T & Guestrin C. 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16: 785–794. https://doi.org/10.1145/2939672.2939785

Chen T, He T, Benesty M, Khotilovich V,Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J, et al (XGBoost contributors). 2023. Package 'xgboost'. Extreme Gradient Boosting. Retrieved from https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

Corley RHV & Tinker PB. 2015. *The Oil Palm*. https://doi.org/10.1002/9781118953297

Dhandapani S, Evers S, Ritz K & Sjцgersten S. 2020. Nutrient and trace element concentrations influence greenhouse gas emissions from Malaysian tropical peatlands. *Soil Use and Management*, 37(1): 138-150. https://doi.org/10.1111/sum.12669

Dhandapani S, Girkin NT & Evers S. 2021. Spatial variability of surface peat properties and carbon emissions in a tropical peatland oil palm monoculture during a dry season. *Soil Use and Management*, 38(1): 381-395. https://doi.org/10.1111/sum.12741

Ditjen Perkebunan. 2011. Sustainable Palm Oil Development Policy. In Indonesia: Kebijakan Pengembangan Kelapa Sawit Berkelanjutan. Seminar on RSPO Implementation in Indonesia. Jakarta, 10 Februari 2011

Efron B. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26. https://doi.org/10.1214/aos/1176344552

Efron B & Tibshirani RJ. 1994. *An Introduction to the Bootstrap*. CRC Press, Boca Raton.

Engels C, Kirkby E, & White P. 2012. Mineral nutrition, yield and source–sink relationships. In Marchner P (Ed.) *Marschner's Mineral Nutrition of Higher Plants*, Academic Press. p.85–133. https://doi.org/10.1016/b978-0-12-384905-2.00005-4

Forkuor G, Hounkpatin OKL, Welp G, & Thiel M. 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PLoS ONE,* 12(1):e0170478. https://doi.org/10.1371/journal.pone.0170478

Friedman JH. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman JH. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378. https://doi.org/10.1016/s0167-9473(01)00065-2

Garson GD. 1991. Interpreting neural network connection weights. *AI Expert*, 6:47-51

Glendining MJ, Dailey AG, Powlson DS, Richter GM, Catt JA & Whitmore AP. 2010. Pedotransfer functions for estimating total soil nitrogen up to the global scale. *European Journal of Soil Science*, 62(1):13–22. https://doi.org/10.1111/j.1365-2389.2010.01336.x

Goh ATC. 1995. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, 9(3):143–151. https://doi.org/10.1016/0954-1810(94)00011-s

Goydaragh MG, Taghizadeh-Mehrjardi R, Jafarzadeh AA, Triantafilis J, & Lado M. 2021. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. *Catena*, 202:105280. https://doi.org/10.1016/j.catena.2021.105280

Greenwell B. 2023. Package 'fastshap'. Fast Approximate Shapley Values. Retrieved from https://cran.r-project.org/web/packages/fastshap/fastshap.pdf

Greenwell B, Boehmke B & Cunningham J. 2022. Package 'gbm'. Generalized Boosted Regression Models. Retrieved from https://cran.rproject.org/web/packages/gbm/index.html

Hall RL, de Santana FB, Grunsky EC, Browne MA, Lowe V, Fitzsimons M, Higgins S, Galagher V & Daly K. 2023. A machine learning approach to predicting plant available phosphorus that accounts for soil heterogeneity and regional variability. *Journal of Soils and Sediments*, https://doi.org/10.1007/s11368-023-03648-y

Harianti M, Sutandi A, Saraswati R, Maswar & Sabiham S. 2018. Enzyme activities in relation to total K, Ca, Mg Fe, Cu, and Zn in the oil palm rhizosphere of Riau's peatlands, Indonesia. *Biotropia,* 3(25):211-223. https://doi.org/10.11598/btb.2018.25.3.862

Hashim SA, Teh CBS & Ahmed OH. 2019. Influence of water table depths, nutrients leaching losses, subsidence of tropical peat soil and oil palm (*Elaeis guineensis* Jacq.) seedling growth. *Malaysian Journal of Soil Science,* 23:13-30.

Ho TK. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844. https://doi.org/10.1109/34.709601

Hounkpatin KO, Bossa AY, Yira Y, Igue MA, & Sinsin BA. 2022. Assessment of the soil fertility status in Benin (West Africa) – Digital soil mapping using machine learning. *Geoderma Regional*, 28:e00444. https://doi.org/10.1016/j.geodrs.2021.e00444

Huang AA & Huang SY. 2023. Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations. *PLOS ONE*, 18(2). https://doi.org/10.1371/journal.pone.0281922

Huang F, Zhang Y, Zhang Y, Nourani V, Li Q, Li L, Shangguan W. 2023. Towards interpreting machine learning models for predicting soil moisture droughts. *Environmental Research Letters*, 18:074002. https://doi.org/10.1088/1748-9326/acdbe0

Hunt J. 2022. Package 'ModelMetrics'. Rapid Calculation of Model Metrics. Retrieved from https://cran.r-project.org/web/packages/ModelMetrics/ModelMetrics.pdf

Iranpanah N, Mansourian A, Tashayo B & Haghighi F. 2010. Spatial semi-parametric bootstrap method for analysis of kriging predictor of random field. *Procedia Environmental Sciences*, 3:81-86. https://doi.org/10.1016/j.proenv.2011.02.015

James G, Witten D, Hastie T, Tibshirani R. 2021. *An Introduction to Statistical Learning: with Applications in R.* New York(US): Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jas K & Dodagoudar G. 2023. Explainable machine learning model for liquefaction potential assessment of soils using XGBoost-SHAP. *Soil Dynamics and Earthquake Engineering*, 165:107662. https://doi.org/10.1016/j.soildyn.2022.107662

Jones EJ, Bishop TF, Malone BP, Hulme PJ, Whelan BM & Filippi P. 2022. Identifying causes of crop yield variability with interpretive machine learning. *Computers and Electronics in Agriculture*, 192:106632. https://doi.org/10.1016/j.compag.2021.106632

Jordan MI & Mitchell TM. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260. https://doi.org/10.1126/science.aaa8415

Karasiak N, Dejoux J-F, Monteil, C & Sheeren D. 2021. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*. https://doi.org/10.1007/s10994-021-05972-1

Karim KH. 2023. Prediction of soil total nitrogen based on total organic carbon using different models in soils from the Iraqi Kurdistan Region. *Passer Journal of Basic and Applied Sciences,* 5(1): 178-182. https://doi.org/10.24271/psr.2023.388581.1274

Kaya F, Keshavarzi A, Francaviglia R, Kaplan G, Basayigit L, & Dedeoğlu M. 2022. Assessing machine learning-based prediction under different agricultural practices for digital mapping of soil organic carbon and available phosphorus. *Agriculture*, 12(7):1062. https://doi.org/10.3390/agriculture12071062

Koh LP, Miettinen J, Liew SC, & Ghazoul J. 2011. Remotely sensed evidence of tropical peatland conversion to oil palm. *Proceedings of the National Academy of Sciences*, 108(12):5127–5132. https://doi.org/10.1073/pnas.1018776108

Kцпцnen, M, Jauhiainen J, Strakovб P, Heinonsalo J, Laiho R, Kusin K, Limin S, & Vasander H. 2018. Deforested and drained tropical peatland sites show poorer peat substrate quality and lower microbial biomass and activity than unmanaged swamp forest. *Soil Biology and Biochemistry*, 123:229-241. https://doi.org/10.1016/j.soilbio.2018.04.028

Kuhn M & Silge J. 2023. *Tidy Modelling with R.* California (US): O'Reilly Media, Inc. https://www.tmwr.org/

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel BR, Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C & Hunt T. 2023.

Package 'caret'. Classification and Regression Training. Retrieved from https://cran.r-project.org/web/packages/caret/index.html

Leifeld J & Menichetti L. 2018. The underappreciated potential of peatlands in global climate change mitigation strategies. *Nature Communications*, 9(1):1071 https://doi.org/10.1038/s41467-018-03406-6

Li X, Zhang C, Behrens H & Holtz F. 2020. Calculating biotite formula from electron microprobe analysis data using a machine learning method based on principal components regression. *Lithos*, 356-357:105371. https://doi.org/10.1016/j.lithos.2020.105371

Lindstromberg S. 2023. The winner's curse and related perils of low statistical power – spelled out and illustrated. *Research Methods in Applied Linguistics*, 2(3):100059. https://doi.org/10.1016/j.rmal.2023.100059

Lipton ZC. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Liu F, Zhang G, Song X, Li D, Zhao Y, Yang J, Wu H, & Yang F. 2020. High-resolution and three-dimensional mapping of soil texture of China. *Geoderma*, 361:114061. https://doi.org/10.1016/j.geoderma.2019.114061

Liu Z, Lei H, Lei L, & Sheng H. 2022. Spatial prediction of total nitrogen in soil surface layer based on machine learning. *Sustainability*, 14:11998. https://doi.org/10.3390/su141911998

Lundberg S & Lee S. 2017. A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. ArXiv. /abs/1705.07874. https://doi.org/10.48550/arXiv.1705.07874

Lyons MB, Keith DA, Phinn SR, Mason TJ & Elith J. 2018. A comparison of resampling methods for remote sensing classification and accuracy assessment. *Remote Sensing of Environment*, 208:145-153. https://doi.org/10.1016/j.rse.2018.02.026

Ma Y, Minasny B, McBratney A, Poggio L, & Fajardo M. 2021. Predicting soil properties in 3D: Should depth be a covariate? *Geoderma*, 383:114794. https://doi.org/10.1016/j.geoderma.2020.114794

McPherron SP, Archer W, Otǒrola-Castillo ER, Torquato MG & Keevil TL. 2022. Machine learning, bootstrapping, null models, and why we are still not 100% sure which bone surface modifications were made by crocodiles. *Journal of Human Evolution*, 164:103071. https://doi.org/10.1016/j.jhevol.2021.103071

Maharana K, Mondal S & Nemade B. 2022. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91-99. https://doi.org/10.1016/j.gltp.2022.04.020

Marschner P & Rengel Z. 2012. Nutrient Availability in Soils. In Marschner P (Ed.) *Marschner's Mineral Nutrition of Higher Plants (Third Edition)*, Academic Press. p.315-330. https://doi.org/10.1016/B978-0-12-384905-2.00012-1

Mashaba-Munghemezulu Z, Chirima GJ, & Munghemezulu C. 2021. Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and Sentinel-2 data. *Sustainability*, 13:11591. https://doi.org/10.3390/su132111591

Mayer M & Stando A. 20203. Package 'shapviz'. SHAP Visualizations. Retrieved from https://cran.r-project.org/web/packages/shapviz/shapviz.pdf

Melling L, Goh KJ, & Hatano R. 2006. Short-term effect of urea on CH_4 flux under the oil palm (*Elaeis guineensis*) on tropical peatland in Sarawak, Malaysia. *Soil Science and Plant Nutrition*, 52(6):788–792. https://doi.org/10.1111/j.1747-0765.2006.00092.x

Mesele SA & Ajiboye GA. 2020. Pedo-transfer functions for predicting total soil nitrogen in different land use types under some tropical environments. *Ghana Journal of Science*, 61 (2):45 – 56. https://dx.doi.org/10.4314/gjs.v61i2.5

Milborrow S. 2023. Package 'earth'. Multivariate Adaptive Regression Splines. Retrieved from https://cran.rproject.org/web/packages/earth/index.html

Moghaddam DD, Rahmati O, Panahi M, Tiefenbacher J, Darabi H, Haghizadeh A, Haghighi AT, Nalivan OA & Tien Bui D. 2020. The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *Catena*, 187:104421. https://doi.org/10.1016/j.catena.2019.104421

Mohidin H, Hanafi MM, Rafii YM, Abdullah SNA, Idris AS, Man S, Idris J, & Sahebi M. 2015. Determination of optimum levels of nitrogen, phosphorus and potassium of oil palm seedlings in solution culture. *Bragantia*, 74(3):247–254. https://doi.org/10.1590/1678-4499.0408

Molnar C. 2023. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/

Nash JF. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1): 48-49. https://doi.org/10.1073/pnas.36.1.48

Noor M, Masganti, Agus F. 2016. Pembentukan dan karakteristik gambut tropika Indonesia. In Agus F, Anda M, Jamil A, Masganti (Eds.). *Lahan Gambut Indonesia: Pembentukan, Karakteristik, dan Potensi Mendukung Ketahanan Pangan*. Bogor (ID): Badan Penelitian dan Pengembangan Pertanian, Badan Penelitian Dan Pengembangan Pertanian. IAARD Press. p.7-15

Olden JD & Jackson DA. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1-2):135–150. https://doi.org/10.1016/s0304-3800(02)00064-9

Padarian J, McBratney AB & Minasny B. 2020. Game theory interpretation of digital soil mapping convolutional neural networks, *SOIL*, 6:389–397. https://doi.org/10.5194/soil-6-389-2020

Padarian J, Morris J, Minasny B, & McBratney AB. 2018. Pedotransfer Functions and Soil Inference Systems. In McBratney AB., Minasny B, Stockmann U (Eds.) *Progress in Soil Science*, Springer Cham. p.195–220. https://doi.org/10.1007/978-3-319-63439-5_7

Padarian J, Minasny B, & McBratney AB. 2020. Machine learning and soil sciences: a review aided by machine learning tools. *Soil*, 6(1):35–52. https://doi.org/10.5194/soil-6-35-2020

Paleckiene R, Navikaite R, & Slinksiene R. 2021. Peat as a raw material for plant nutrients and humic substances. *Sustainability*, 13(11):6354. https://doi.org/10.3390/su13116354

Parsaie F, Farrokhian Firouzi A, Mousavi SR, Rahmani A, Sedri MH, & Homaee M. 2021. Large-scale digital mapping of topsoil total nitrogen using machine learning models and associated uncertainty map. *Environmental Monitoring and Assessment*, 193(4):162. https://doi.org/10.1007/s10661-021-08947-w

Pijlman J, Holshof G, van den Berg W, Ros GH, Erisman JW & van Eekeren N. 2020. Soil nitrogen supply of peat grasslands estimated by degree days and soil organic matter content. *Nutrient Cycling in Agroecosystems*, 117. https://doi.org/10.1007/s10705-020-10071-z

Pulunggono HB. 2019. Nutrient Dynamic on Peatland in Oil Palm Plantation. Dissertation. Bogor (ID): IPB University

Pulunggono HB, Anwar S, Mulyanto B, Sabiham S. 2016. Dynamics and distribution of peat water macro nutrients (N, P, K, Ca, Mg and S) in oil palm plantation based on season, peat thickness, chanel distance and plant age. 15th International Peat Congress, Kuching, Malaysia 2016. A-263, International Peat Society. pp.58-61

Pulunggono HB, Zulfajrin M, Hartono A. 2020. Selected chemical peat properties distribution in palm oil plantation and its relationship with depth layer and distance from mineral soil derived from ultrabasic rocks. *Journal of Soil Science and Environment*, 22(1): 22-28. https://doi.org/10.29244/jitl.22.1.22-28

Pulunggono HB, Zulfajrin M, Irsan F. 2021. Distribution of nickel (Ni) in peatland situated alongside mineral soil derived from ultrabasic rocks. *SAINS TANAH – Journal of Soil Science and Agroclimatology*, 18(1):15-26. https://dx.doi.org/10.20961/stjssa.v18i1.45417

Pulunggono HB, Madani YA, Zulfajrin M, & Yusrizal. 2022a. Identifying the underlying factors and variables governing macronutrients in cultivated tropical peatland using regression tree approach. *CELEBES Agricultural*, 3(1):43–61. https://doi.org/10.52045/jca.v3i1.353

Pulunggono HB, Nurazizah LL, Anwar S, Sabiham S. 2022b. Assessing the distribution of total Fe, Cu, and Zn in tropical peat at an oil palm plantation and their relationship with several environmental factors. *Journal of Degraded and Mining Lands Management*, 9(2): 3349-3358. https://doi.org/10.15243/jdmlm.2022.092.3349

Pulunggono HB, Kartika VW, Nadalia D, Nurazizah LL, Zulfajrin M. 2022c. Evaluating the changes of Ultisol chemical properties and fertility characteristics due to animal manure amelioration. *Journal of Degraded and Mining Lands Management*, 9(3):3545-3560. https://doi.org/10.15243/jdmlm.2022.093.3545

Pulunggono HB, Hanifah N, Nadalia D, Zulfajrin M, Nurazizah LL, Mubarok H, Tambusai N, Anwar S, Sabiham S. 2022d. Declined peat heterotrophic respiration as consequences from zeolite amendment simulation: coupling descriptive and predictive modelling approaches. *Journal of Degraded and Mining Lands Management*, 10(1):3889-3904. https://doi.org/10.15243/jdmlm.2022.101.3889

Pulunggono HB, Fitriana S, Nadalia D, Zulfajrin M, Nurazizah LL, Mubarok H, Tambusai N, Anwar S, Sabiham S. 2022e. Simulating and modeling CO_2 flux emitted from decomposed oil palm root cultivated at tropical peatland as affected by water content and residence time. *Journal of Degraded and Mining Lands Management*, 9(4):3663-3676. https://doi.org/10.15243/jdmlm.2022.094.3663

Pulunggono HB, Siswanto, Mubarok H, Widiastuti H, Tambusai N, Zulfajrin M, Anwar S, Taniwiryono D, Sumawinata B & Sabiham. S. 2022. Seasonal litter contribution to total peat

respiration from drained tropical peat under mature oil palm plantation. *Journal of Degraded and Mining Lands Management,* 9(2): 3247-3263, https://doi.org/10.15243/jdmlm.2022.092.3247

Quinlan R. 1992. Learning with Continuous Classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp.343-348.

Quinlan R. 1993. Combining Instance-Based and Model-Based Learning. Proceedings of the Tenth International Conference on Machine Learning, pp.236-243.

Rajput D, Wang WJ & Chen CC. 2023. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 24(48):1-17. https://doi.org/10.1186/s12859-023-05156-9

Rasmusen E. 1989. *Games and information: An introduction to game theory.* Oxford (UK): Blackwell

Ribeiro K, Pacheco FS, Ferreira JW, Sousa-Neto ER, Hastie A, Krieger Filho GC, Alvalб PC, Forti MC, & Ometto JP. 2020. Tropical peatlands and their contribution to the global carbon cycle and climate change. *Global Change Biology*, 27(3):489–505. https://doi.org/10.1111/gcb.15408

Ribeiro MT, Singh S & Guestrin C. 2016. Model-agnostic interpretability of machine learning. 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NYArXiv. /abs/1606.05386. https://doi.org/10.48550/arXiv.1606.05386

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schrцder B, Thuiller W, Warton DI, Wintle BA, Hartig F & Dormann CF. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913-929. https://doi.org/10.1111/ecog.02881

Rossiter DG. 2018. Past, present & future of information technology in pedometrics. *Geoderma*, 324:131-137. https://doi.org/10.1016/j.geoderma.2018.03.009

Rьckauf U, Augustin J, Russow R & Merbach W. 2004. Nitrate removal from drained and reflooded fen soils affected by soil N transformation processes and plant uptake. *Soil Biology and Biochemistry*, 36(1):77-90. https://doi.org/10.1016/j.soilbio.2003.08.021

Sakhaee A, Gebauer A, LieЯ M, & Don A. 2022. Spatial prediction of organic carbon in German agricultural topsoil using machine learning algorithms, *soil*, 8:587–604. https://doi.org/10.5194/soil-8-587-2022

Sangok FE, Maie N, Melling L & Watanabe A. 2017. Evaluation on the decomposability of tropical forest peat soils after conversion to an oil palm plantation. *Science of The Total Environment*, 587-588: 381–388. https://doi.org/10.1016/j.scitotenv.2017.02.165

Sarkar SK, Rudra RR, Nur MS & Das PC. 2023. Partial least-squares regression for soil salinity mapping in Bangladesh. *Ecological Indicators*, 154:110825. https://doi.org/10.1016/j.ecolind.2023.110825

Scharlemann JP, Tanner EV, Hiederer R, & Kapos V. 2014. Global soil carbon: understanding and managing the largest terrestrial carbon pool. *Carbon Management*, 5(1):81–91. https://doi.org/10.4155/cmt.13.77

Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg A, Crowley J. 2023. Package 'Ggally'. Extension to 'ggplot2. Retrieved from https://cran.r-project.org/web/packages/GGally/GGally.pdf

Shapley LS. 1953. *A Value for n-Person Games, Vol II of Contributions to the Theory of Games*. Princeton (US): Princeton University Press

Shi S, Hou M, Gu Z, Jiang C, Zhang W, Hou M, Li C, & Xi Z. 2022. Estimation of heavy metal content in soil based on machine learning models. *Land*, 11:1037. https://doi.org/10.3390/land11071037

Shi Y, Zhang X, Wang Z, Xu Z, He C, Sheng L, Liu H, & Wang Z. 2021. Shift in nitrogen transformation in peatland soil by nitrogen inputs. *Science of The Total Environment*, 764:142924. https://doi.org/10.1016/j.scitotenv.2020.142924

Sihag P, Keshavarzi A, & Kumar V. 2019. Comparison of different approaches for modeling of heavy metal estimations. *SN Applied Sciences*, 1(7):780. https://doi.org/10.1007/s42452-019-0816-6

Soil Survey Staff. 2014. *Key to Soil Taxonomy*. *Twelfth Edition*. Washington DC (US): Natural Resources Conservation Services, United States Department of Agriculture.

Song X, Rossiter DG, Liu F, Wu H, Zhao X, Cao Q, & Zhang G. 2020. Can pedotransfer functions based on environmental variables improve soil total nutrient mapping at a regional scale?. *Soil and Tillage Research*, 202:104672. https://doi.org/10.1016/j.still.2020.104672

trumbelj E & Kononenko I. 2013. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3): 647–665. https://doi.org/10.1007/s10115-013-0679-x

Subardja D, Ritung S, Anda M, Sukarman, Suryani E, & Subandiono RE. 2014. *Petunjuk Teknis Klasifikasi Tanah Nasional*. Bogor (ID):Balai Besar Penelitian dan Pengembangan Sumberdaya Lahan Pertanian, Badan Penelitian dan Pengembangan Pertanian.

Sui Q-R, Chen Q-H, Wang D-D. Tao Z-G. 2023. Application of machine learning to the Vs-based soil liquefaction potential assessment. *Journal of Mountain Science*, 20:2197–2213. https://doi.org/10.1007/s11629-022-7809-4

Tang L, Schucany W, Woodward W & Gunst R. 2006. *A parametric spatial bootstrap*. Technical Report SMU-TR-337. Dallas, Southern Methodist University

Tran VQ. 2022. Predicting and investigating the permeability coefficient of soil with aided single machine learning algorithm. *Complexity*, 8089428. https://doi.org/10.1155/2022/8089428

Trontelj ml. J & Chambers O. 2021. Machine learning strategy for soil nutrients prediction using spectroscopic method. *Sensors*, 2:4208. https://doi.org/10.3390/s21124208

von Uexkьll HR & Fairhurst TH. 1999. Some nutritional disorders in oil palm. *Better Crops International*, 13(1):16-21.

Wadoux C. 2023. Interpretable spectroscopic modelling of soil with machine learning. *European Journal of Soil Science*, 74(3):e13370. https://doi.org/10.1111/ejss.13370

Wadoux AMJ-C, Romбn-Dobarco M, & McBratney AB. 2020a. Perspectives on data-driven soil research. *European Journal of Soil Science*, 72(4):1675–1689. https://doi.org/10.1111/ejss.13071

Wadoux AMJ-C, Minasny B, & McBratney AB. 2020b. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, 210:103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wadoux AMJ-C & Molnar C. 2022. Beyond prediction: methods for interpreting complex models of soil variation. *Geoderma*, 422:115953. https://doi.org/10.1016/j.geoderma.2022.115953

Wadoux AMJ-C, Odeh IO, & McBratney AB. 2021. *Overview of Pedometrics*. Reference Module in Earth Systems and Environmental Sciences. https://doi.org/10.1016/B978-0-12-822974-3.00001-X

Wadoux C, Samuel-Rosa A, Poggio L & Mulder VL. 2020c. A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science*, 71(2):133-136. https://doi.org/10.1111/ejss.12909

Warren M, Hergoualc'h K, Kauffman JB, Murdiyarso D, & Kolka R. 2017. An appraisal of Indonesia's immense peat carbon stock using national peatland maps: uncertainties and potential losses from conversion. *Carbon Balance and Management*, 12(1):12. https://doi.org/10.1186/s13021-017-0080-2

Wen L & Hughes M. 2020. Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. *Remote Sensing*, 12(10):1683. https://doi.org/10.3390/rs12101683

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, Franзois R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Mьller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K & Yutani H. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43): 1686. https://doi.org/10.21105/joss.01686

Wright MN, Wager S & Probst P. 2023. Package 'ranger'. A Fast Implementation of Random Forests. Retrieved from https://cran.rproject.org/web/packages/ranger/index.html

Xu Y & Goodacre R. 2018. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*. 2:249–262. https://doi.org/10.1007/s41664-018-0068-2

Xu S, Wang M, Shi X, Yu Q, & Zhang Z. 2021. Integrating hyperspectral imaging with machine learning techniques for the high-resolution mapping of soil nitrogen fractions in soil profiles. *Science of The Total Environment*, 754:142135. https://doi.org/10.1016/j.scitotenv.2020.142135

Xu S, Zhao Y, Wang M & Shi X. 2018. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. *Geoderma*, 310:29-43. https://doi.org/10.1016/j.geoderma.2017.09.013

Young GA. 1994. Bootstrap: More than a stab in the dark? *Statistical Science.* 9(3): 382-395. https://doi.org/10.1214/ss/1177010383

Yu Z, Loisel J, Brosseau DP, Beilman DW, & Hunt SJ. 2010. Global peatland dynamics since the Last Glacial Maximum. *Geophysical Research Letters*, 37(13):L13402. https://doi.org/10.1029/2010gl043584

Zhao SL, Gupta SC, Huggins DR & Moncrief JF. 2001. Tillage and nutrient source effects on surface and subsurface water quality at corn planting. *Journal of Environmental Quality*, 30(3):998-1008. https://doi.org/10.2134/jeq2001.303998x