Ottimizzazione degli iperparametri nella stima di volume e carbonio forestale Accuratezza e costo computazionale

G. D'Amico[1][0000-0002-2341-3268], A. Montaghi[2][0000-0001-8351-0230], M. Niccoli[1][0009-0005-9744-1295], E. Vangi[1][0000-0002-9772-2258], C. Borghi[1][0000-0002-5334-7548], G. Chirici[1][0000-0002-0669-5726]

[1] Department of Agriculture, Food, Environment and Forestry (DAGRI), University of Florence, Florence, Italy; giovanni.damico@unifi.it, mattia.niccoli@unifi.it, elia.vangi@unifi.it, costanza.borghi@unifi.it, gherardo.chirici@unifi.it

[2] Institute for BioEconomy (IBE), National Research Council (CNR), Sesto Fiorentino, Italy. alessandro.montaghi@cnr.it

Abstract. Su plot inventariali e dati multisorgente confrontiamo cinque modelli per volume e carbonio forestale, due strategie di ottimizzazione e tre budget. Extreme Gradient Boosting (XGBoost) beneficia maggiormente del tuning. Nei modelli di boosting, da 100 a 1000 tentativi il guadagno mediano nell'errore relativo è di 0,09 punti percentuali, con tempi circa dieci volte maggiori. Budget moderati risultano adeguati in molte configurazioni.

Parole chiave: Inventario forestale, telerilevamento, machine learning, iperparametri, costo computazionale.

1 Introduzione

Gli inventari forestali forniscono informazioni essenziali sulla consistenza delle risorse legnose e sui serbatoi di carbonio [1]. L'integrazione dei rilievi a terra con le osservazioni satellitari consente di stimare le variabili forestali anche al di fuori dei punti campionati. I modelli di machine learning apprendono queste relazioni dai dati, combinando predittori telerilevati con informazioni ambientali e strutturali. Gli ensemble di alberi decisionali combinano le predizioni di più alberi per rappresentare relazioni non lineari e interazioni tra predittori. Queste caratteristiche ne hanno favorito l'ampia diffusione nella stima spazialmente esplicita di variabili forestali mediante dati telerilevati.

Nelle applicazioni di telerilevamento forestale, gli ensemble di alberi decisionali sono tra gli approcci di machine learning più utilizzati per stimare variabili quantitative. Random Forest (RF) si è affermato come metodo di riferimento, spesso adottato come scelta iniziale nella costruzione dei modelli [2, 3]. RF combina le predizioni di alberi addestrati su campioni bootstrap, ottenuti mediante estrazione con reinserimento delle osservazioni dal set di addestramento. Durante la costruzione di ciascun albero, un iperparametro determina il numero di predittori candidati a ogni suddivisione; le predizioni dei singoli alberi vengono quindi mediate per ottenere la stima finale [3].

Accanto a RF, trovano crescente impiego XGBoost (Extreme Gradient Boosting) e LightGBM (Light Gradient Boosting Machine), entrambi basati sul gradient boosting [4, 5]. In questo approccio, gli alberi sono costruiti sequenzialmente: ogni nuovo albero contribuisce a correggere gli errori dell'ensemble corrente, seguendo il gradiente della funzione di perdita. Questa modalità di apprendimento rende particolarmente rilevante la scelta degli iperparametri che regolano il contributo e la complessità degli alberi.

Per LightGBM, questo studio considera tre configurazioni: Gradient Boosting Decision Tree (GBDT), che impiega il boosting tradizionale; Gradient-based One-Side Sampling (GOSS), che mantiene le osservazioni con gradienti elevati e campiona una parte di quelle con gradienti più piccoli; e Dropouts meet Multiple Additive Regression Trees (DART), che esclude temporaneamente alcuni alberi durante l'apprendimento per limitarne l'eccessiva specializzazione [5, 6]. A seconda dell'algoritmo, iperparametri quali il numero di alberi, la profondità massima e il tasso di apprendimento controllano la struttura del modello e il processo di addestramento. Essi sono definiti prima dell'apprendimento e si distinguono dai parametri interni, stimati dai dati durante il training [7].

L'ottimizzazione degli iperparametri (Hyperparameter Optimization, HPO), o tuning, mira a ottenere prestazioni migliori in fase di validazione. Random Search campiona configurazioni dalle distribuzioni definite, senza utilizzare i risultati precedenti [8]. Optuna è un framework che, con il campionatore Tree-structured Parzen Estimator (TPE), adatta la ricerca alle configurazioni già valutate [9]. Il beneficio dipende dall'algoritmo e dal problema [2]. Questo contributo confronta modelli predefiniti e ottimizzati per la stima di volume e carbonio delle foreste italiane, valutando l'accuratezza, il tempo di calcolo e il guadagno marginale da 100 a 1000 tentativi.

2 Materiali e metodi

I dati provengono dal terzo Inventario Nazionale delle Foreste e dei Serbatoi Forestali di Carbonio, riferito al 2015 (INFC2015) [1]. In particolare, sono stati utilizzati i 6.993 plot della terza fase, dedicata ai rilievi quantitativi. Le variabili di risposta sono il volume legnoso (Growing Stock Volume, GSV, $m^3\ ha^{-1}$) e lo stock di carbonio (Carbon Stock, CS, Mg C ha^{-1}).

I predittori sono stati organizzati in quattro gruppi: i. Landsat e ii. Sentinel (Sentinel-2), basati sui rispettivi dati ottici; iii. MIX, una combinazione multisensore selezionata; iv. MIX_small, una configurazione con meno predittori (Tab. 1). Il nucleo comprende bande, indici spettrali, componenti Tasseled Cap e variabili derivate dal modello digitale di elevazione (Digital Elevation Model, DEM) Copernicus. Sentinel-1 fornisce retrodiffusione radar ad apertura sintetica (Synthetic Aperture Radar, SAR); VH indica trasmissione verticale e ricezione orizzontale. I dati ottici Landsat e Sentinel-2 derivano da compositi annuali sviluppati con l'approccio medoid [10].

Gli indici comprendono Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Normalized Burn Ratio (NBR), Normalized Difference Water Index (NDWI), Normalized Mangrove Vegetation Index (NMVI),

Normalized Difference Built-up Index (NDBI), Normalized Difference Moisture Index (NDMI) e Moisture Stress Index (MSI). Le componenti Tasseled Cap rappresentano la luminosità, il verde della vegetazione e l'umidità spettrale e, per Landsat, una componente angolare; TCG indica Tasseled Cap Greenness. I prefissi LT e S2 identificano Landsat e Sentinel-2; SWIR2 è la seconda banda dell'infrarosso a onde corte (Short-Wave Infrared, SWIR). Oltre agli indici spettrali, è utilizzata anche la distanza euclidea (Euclidean Distance, ED) associata alla selezione medoid, quale indicatore della rappresentatività spettrale dell'osservazione.

Tutti i gruppi includono covariate climatiche di radiazione, precipitazione, temperatura e deficit di pressione di vapore, e un modello di altezza della copertura arborea (Canopy Height Model, CHM) di Tolan et al. [11]. Landsat, Sentinel e MIX includono anche mappe pregresse di volume e carbonio, escluse da MIX_small. I conteggi della Tab. 1 riguardano solo i nuclei ottico, radar e topografico, non il totale dei predittori finali. Il confronto modifica quindi sia la quantità sia il contenuto informativo delle variabili.

Tab. 1. Nucleo ottico, radar e topografico. N indica il numero di queste variabili, escluse integrazioni climatiche, altezza della copertura e mappe pregresse. Le variabili risposta e la categoria forestale non rientrano nel conteggio.

Insieme	N	Composizione
Landsat	19	Quota, pendenza ed esposizione; 6 bande ottiche; 6 indici spettrali: NDVI, EVI, NBR, NDWI, EMVI, NDBI; 4 componenti Tasseled Cap.
Sentinel	17	Quota, pendenza ed esposizione; 5 bande ottiche; 5 indici (MSI, NDMI, NDVI, EVI, NBR); 3 componenti Tasseled Cap; S2_ED.
MIX	14	Quota e pendenza; 4 bande ottiche; 4 indici spettrali; 3 componenti Tasseled Cap; retrodiffusione VH Sentinel-1 in orbita ascendente.
MIX_small	8	Quota; LT_SWIR2, LT_NBR e LT_TCG; S2_SWIR2 e S2_NBR; retrodiffusione VH Sentinel-1 in orbite ascendente e discendente.

Il training (80%) serve all'addestramento e alla selezione dei modelli; il test (20%) è riservato alla valutazione finale. La partizione casuale è stratificata per categoria forestale, identificata dal campo CFOR del database, per conservarne approssimativamente le proporzioni. Gli stessi plot sono assegnati ai due insiemi per tutti i gruppi di predittori. Si confrontano cinque configurazioni: RF, XGBoost con booster gbtree (boosting basato su alberi) e LightGBM nelle configurazioni GBDT, GOSS e DART.

Per ogni combinazione di modello, predittori e variabile risposta si stima una baseline con le impostazioni predefinite del software, indicate nelle figure come Default. Seguono Random Search, implementata con RandomizedSearchCV, e Optuna con TPE, con budget di 10, 100 e 1000 tentativi. Ogni tentativo valuta una configurazione di iperparametri: questo budget è distinto dal numero di alberi o dalle iterazioni di apprendimento del modello.

L'HPO utilizza la cross-validation a cinque partizioni (fold) del training: a turno, quattro partizioni vengono utilizzate per l'addestramento e la quinta per la validazione. Le prestazioni guidano la selezione; il modello scelto è riaddestrato sull'intero training

e valutato sul test, esterno all'HPO. Il disegno comprende 240 procedure di ottimizzazione (5 modelli × 4 gruppi × 2 target × 2 strategie × 3 budget), ciascuna con un modello selezionato, e 40 modelli predefiniti.

Per la valutazione dell'accuratezza sono stati usati il Root Mean Square Error (RMSE), cioè la radice della media dei quadrati delle differenze tra osservazioni e predizioni, e il RMSE%, ottenuto dividendo RMSE per la media osservata del target nel test e moltiplicando il rapporto per 100.

L'effetto del tuning è espresso come ΔRMSE% = RMSE%ottimizzato − RMSE%predefinito: valori negativi indicano miglioramenti. Il guadagno marginale è $G_{100\rightarrow1000}$ = $RMSE\%_{100}$ − $RMSE\%_{1000}$: valori positivi indicano miglioramenti da 100 a 1000 tentativi. Si confrontano gli stessi target, predittori, algoritmi e strategie. Il rapporto t_{1000}/t_{100} quantifica l'aumento dei tempi. Le differenze di RMSE% sono espresse in punti percentuali, abbreviati nelle figure come p.p. I minimi tra strategie e budget restano riepiloghi descrittivi ex post sul test.

Nelle figure di costo e beneficio, tempo e ΔRMSE% sono sintetizzati con mediane separate sui quattro gruppi di predittori, per ciascuna combinazione di algoritmo, target, strategia e budget. Il punto mediano non è necessariamente una singola esecuzione. Le sintesi descrivono configurazioni diverse, non ripetizioni indipendenti della stessa procedura.

3 Risultati

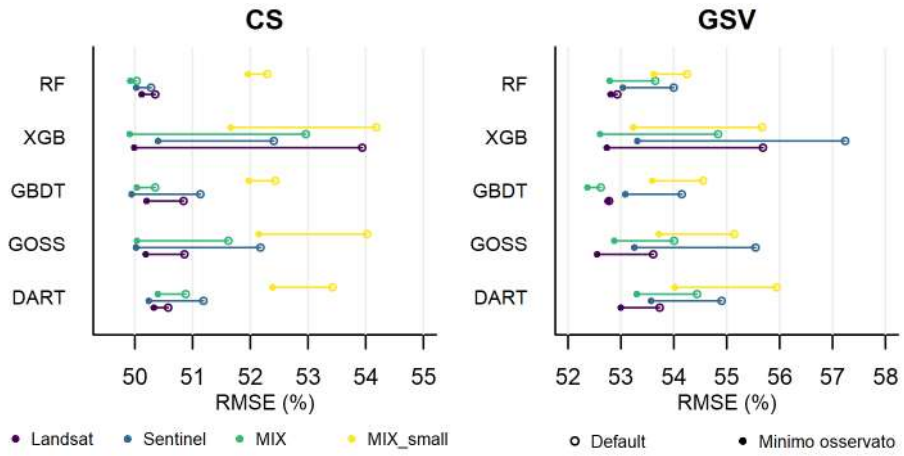

Fig. 1. RMSE% dei modelli predefiniti e minimo osservato tra strategie e budget di HPO. CS indica lo stock di carbonio e GSV il volume legnoso; XGB: XGBoost. GBDT, GOSS e DART indicano le configurazioni LightGBM descritte nell'introduzione. I colori distinguono i quattro gruppi di predittori. Il minimo è una sintesi descrittiva ex post delle prestazioni sul test.

Considerando le 20 combinazioni di algoritmo e predittori per ciascuna variabile risposta, la mediana di RMSE% dei modelli predefiniti è 51,40% per CS e 54,34% per GSV. Se per ciascuna combinazione si considera il minimo osservato tra le sei

procedure di HPO, le mediane sono rispettivamente del 50,19% e del 53,06%. XGBoost mostra il beneficio maggiore: la riduzione mediana rispetto alla baseline, calcolata sui quattro gruppi di predittori, è di 2,79 punti percentuali per CS e 2,69 per GSV. RF presenta riduzioni più contenute, pari a 0,24 e 0,75 punti; le configurazioni di LightGBM presentano risposte intermedie (Fig. 1).
Il beneficio del minimo osservato non implica che tutte le procedure di tuning migliorino la baseline. Alcune configurazioni di RF, infatti, peggiorano fino a 6,57 punti di RMSE% per CS e 9,93 per GSV, soprattutto con budget ridotti. Anche LightGBM-DART presenta configurazioni sfavorevoli. In XGBoost gran parte del recupero rispetto alle impostazioni predefinite si osserva già con 10 tentativi, mentre i budget maggiori determinano variazioni più contenute. Il confronto non identifica una strategia di ricerca superiore in tutte le combinazioni di algoritmo e predittori. La Fig. 2 collega questi risultati ai tempi di ottimizzazione.
Le differenze tra gruppi di predittori persistono dopo il tuning. Per CS, i minimi osservati dei modelli basati su Landsat, Sentinel e MIX ricadono complessivamente nell'intervallo 49,91–50,40% di RMSE%, mentre quelli di MIX_small variano tra 51,66 e 52,38%. Per GSV gli intervalli sono rispettivamente 52,37–53,57% e 53,23–54,02%. Il gruppo con meno predittori risulta dunque generalmente meno accurato, soprattutto per CS; l'HPO non annulla il divario legato alla diversa composizione dei dati.

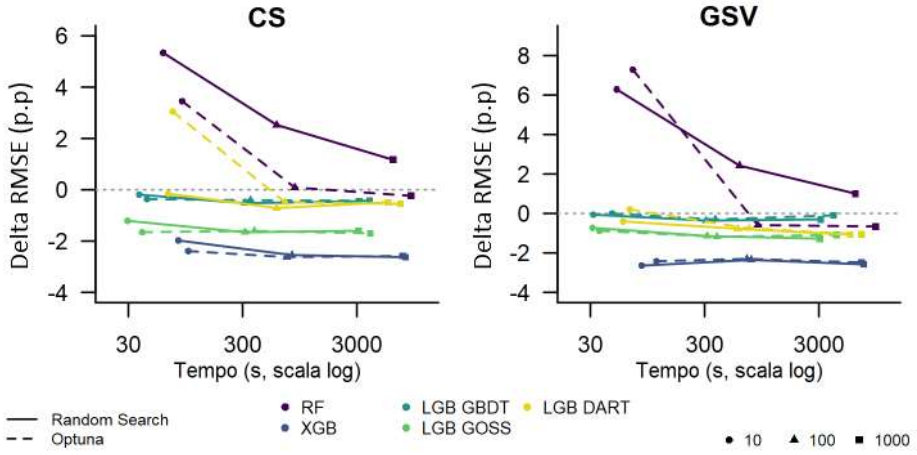

Fig. 2. Tempo di HPO e ΔRMSE% rispetto ai modelli predefiniti. Ogni punto riporta mediane separate sui quattro gruppi, per target, algoritmo, strategia e budget. Valori negativi indicano miglioramenti. XGB e LGB abbreviano XGBoost e LightGBM; p.p. indica punti percentuali. Le linee collegano budget diversi, senza implicare una ricerca cumulativa.

Il confronto tra 100 e 1000 tentativi comprende 64 coppie per i quattro modelli di boosting (4 modelli × 4 gruppi di predittori × 2 target × 2 strategie). Il guadagno mediano è di 0,089 punti di RMSE%, mentre il rapporto mediano dei tempi è 10,05. In 60 delle 64 coppie la variazione assoluta di errore non supera 0,5 punti percentuali, includendo sia miglioramenti sia peggioramenti. RF costituisce il caso con il maggior

recupero a budget elevati, soprattutto con Random Search: il guadagno mediano sui quattro gruppi è di 1,69 punti per CS e 1,42 per GSV. Con Optuna i corrispondenti valori sono 0,27 e 0,20 punti (Fig. 3).

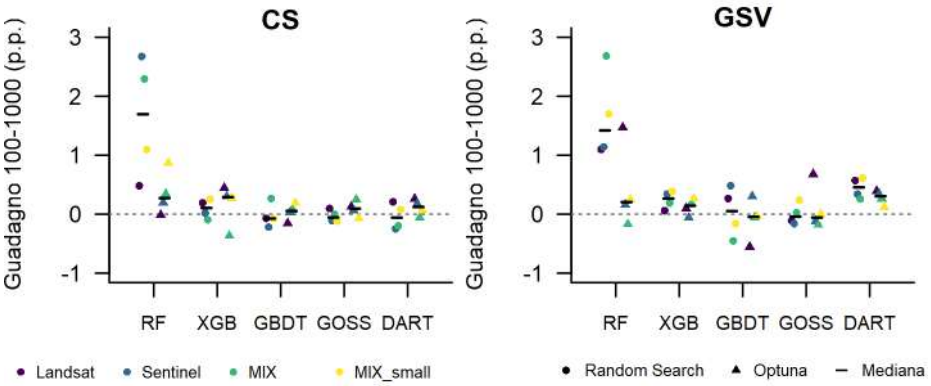

Fig. 3. Guadagno marginale da 100 a 1000 tentativi, calcolato come $RMSE\%_{100} - RMSE\%_{1000}$. I colori distinguono i gruppi di predittori e i simboli le strategie; i segmenti neri rappresentano, per ciascuna strategia, la mediana tra i quattro gruppi. Valori positivi indicano miglioramenti, negativi peggioramenti. Le differenze sono espresse in punti percentuali.

4 Discussione

Il beneficio del tuning deve essere valutato in relazione alla qualità della baseline. XGBoost mostra margini di miglioramento elevati e raggiunge prestazioni prossime a quelle dei modelli già competitivi con le impostazioni predefinite, in accordo con la sensibilità agli iperparametri documentata in letteratura [7, 12]. In RF, il peggioramento di alcune configurazioni è compatibile con limiti dello spazio di ricerca che riducono eccessivamente la complessità degli alberi. Questa interpretazione riguarda gli intervalli esplorati e non dimostra che l'ottimizzazione di RF sia intrinsecamente svantaggiosa. È quindi opportuno mantenere il modello predefinito come termine di confronto durante la selezione in fase di validazione.

Nei modelli di boosting, l'aumento del budget da 100 a 1000 tentativi richiede tempi circa dieci volte maggiori, ma produce variazioni dell'errore generalmente contenute. Il recupero più marcato di RF con Random Search va interpretato in relazione alle prestazioni della baseline: migliorare una configurazione meno accurata ma ottimizzata non implica necessariamente superare il modello predefinito. I risultati sostengono budget moderati come punto di partenza, da estendere qualora la validazione mostri benefici ancora rilevanti. Non definiscono tuttavia né una soglia di arresto universale né una strategia di ricerca sempre preferibile.

Il confronto MIX–MIX_small modifica sia il numero sia il tipo dei predittori. Oltre a escludere le mappe pregresse, MIX_small include la componente Tasseled Cap Greenness di Landsat e la retrodiffusione discendente di Sentinel-1, assenti dal nucleo di MIX. Non rappresenta quindi un sottoinsieme stretto con cui isolare il solo effetto

della dimensionalità. Analogamente, la riduzione della dispersione tra i modelli non dimostra stabilità né tra le esecuzioni né nello spazio. Le mediane riportate descrivono le configurazioni disponibili e non costituiscono intervalli di confidenza.

Lo studio riguarda il contesto forestale italiano e utilizza una singola partizione casuale dei plot, stratificata per categoria forestale. Il test, mantenuto esterno all'ottimizzazione, valuta le prestazioni su osservazioni escluse dall'addestramento nello stesso dominio geografico. La trasferibilità ad altre aree richiede ulteriori verifiche, mediante partizioni spaziali che separino geograficamente il training e il test, e mediante applicazioni in contesti forestali differenti.

La selezione ex post del minimo sul test, utilizzata nelle figure e nei risultati, per ciascuna combinazione di algoritmo, predittori e variabile di risposta, consente di descrivere il beneficio massimo osservato del tuning; la conferma della configurazione più efficace richiede ulteriori partizioni dei dati o un insieme di valutazione indipendente.

5 Conclusioni

L'HPO può migliorare la stima di volume e carbonio forestale, ma il beneficio dipende dall'algoritmo, dalla baseline e dallo spazio di ricerca. Nel caso esaminato, XGBoost mostra il recupero più consistente, mentre RF è già competitivo con le impostazioni predefinite. Nei quattro modelli di boosting, il passaggio da 100 a 1000 tentativi produce un guadagno mediano di circa 0,09 punti di RMSE% con tempi circa dieci volte maggiori. Per applicazioni analoghe, un approccio operativo consiste nel partire da budget moderati, confrontare le prestazioni con la baseline e ampliare la ricerca in base ai risultati di validazione. La trasferibilità di queste indicazioni richiede partizioni spaziali, ripetizioni indipendenti dell'HPO e confronti in contesti forestali diversi.

Riferimenti bibliografici

[1] Gasparini, P., Di Cosmo, L., Floris, A., De Laurentis, D. (a cura di): Italian National Forest Inventory—Methods and Results of the Third Survey. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98678-0

[2] Freeman, E.A., Moisen, G.G., Coulston, J.W., Wilson, B.T.: Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research 46(3), 323–339 (2016). https://doi.org/10.1139/cjfr-2014-0562

[3] Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[4] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[5] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30, 3146–3154 (2017).

[6] Korlakai Vinayak, R., Gilad-Bachrach, R.: DART: Dropouts meet Multiple Additive Regression Trees. Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics 38, 489–497 (2015).
[7] Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., Lindauer, M.: Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Mining and Knowledge Discovery 13(2), e1484 (2023). https://doi.org/10.1002/widm.1484
[8] Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research 13, 281–305 (2012).
[9] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701
[10] D'Amico, G., et al.: A multi-source remote sensing dataset for large-scale forest monitoring. Data in Brief 67, 112945 (2026). https://doi.org/10.1016/j.dib.2026.112945
[11] Tolan, J., et al.: Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. Remote Sensing of Environment 300, 113888 (2024). https://doi.org/10.1016/j.rse.2023.113888
[12] Probst, P., Boulesteix, A.-L., Bischl, B.: Tunability: Importance of hyperparameters of machine learning algorithms. Journal of Machine Learning Research 20(53), 1–32 (2019).