REMOTE SENSING LETTERS

Forest signal detection for photon counting LiDAR using Random Forest

Bowei Chen[a,b], Yong Pang[b*], Zengyuan Li[b], Hao Lu[c], Peter North[d], Jacqueline Rosette[d] and Min Yan[a]

[a]Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China; [b]Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China; [c]School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; [d]Global Environmental Modelling and Earth Observation (GEMEO), Department of Geography, Swansea University, Swansea, SA2 8PP, United Kingdom

ARTICLE HISTORY
Compiled October 13, 2019

ABSTRACT
ICESat (The Ice, Cloud, and Land Elevation Satellite)-2, as the new generation of NASA (National Aeronautics and Space Administration)'s ICESat mission, had been successfully launched in September 2018. The sensor onboard the satellite is a newly designed photon counting LiDAR (Light Detection And Ranging) system for the first time used in space. From the currently released airborne simulation data, it can be seen that there exist numerous noise photons scattering from the atmosphere to even below the ground, especially for the vegetation areas. Therefore, relevant research on methods to distinguish the signal photons effectively is crucial for further forestry applications. In this paper, a machine learning based approach was proposed to detect the potential signal photons from 14 MATLAS datasets across 3 sites in the USA. We first chose 3 representative and stable features from the 12 statistical features to train and build the Random Forest classifier, then we quantitatively investigated the accuracy, the factors which influence the accuracy and the model transferability across different sites. We found that k-NN (k-Nearest Neighbour) distance and the reachability of the photon towards the nearby signal centre showed good stability and contributed to a robust model establishment. The relevant quantitative assessment demonstrated that the machine learning approach could achieve high detection accuracy over 85% based on a very limited number of samples even in rough terrain conditions. Further analysis proved the potential of model transferability across different sites. These findings indicated that our methods would be of use for future studies of ICESat-2 data for vegetation applications.

KEYWORDS
Photon counting LiDAR; ICESat-2; machine learning; classification

1. Introduction

The first generation of NASA (National Aeronautics and Space Administration)'s ICESat (The Ice, Cloud, and Land Elevation Satellite) mission (Yu et al. 2010) showed many successful applications in mapping important forest parameters such as tree

CONTACT Yong Pang. Email: pangy@ifrit.ac.cn

height and biomass at large scale using the spaceborne LiDAR (Light Detection And Ranging) system (Lefsky et al. 2005; Duncanson, Niemann, and Wulder 2010; Los et al. 2012; Rosette et al. 2015). ICESat-2, a successor to the ICESat mission, has been successfully launched in September 2018 for the data continuity of the spaceborne LiDAR. In contrast to the GLAS (Geoscience Laser Altimeter System) waveform system onboard the ICESat, ICESat-2 has adopted a newly designed system named ATLAS (Advanced Topographic Laser Altimeter System), which is a micro-pulse, multi-beam photon counting LiDAR system working at a wavelength of 532 nm (Evans 2014). The capability of the ATLAS has been pre-validated using four types of airborne simulation data named SIMPL (the Slope Imaging Multi-polarisation Photon-counting Lidar), MABEL (the Multiple Altimeter Beam Experimental Lidar), MATLAS and SIGMA SPACE SPL Prototype during the past few years (Markus et al. 2017).

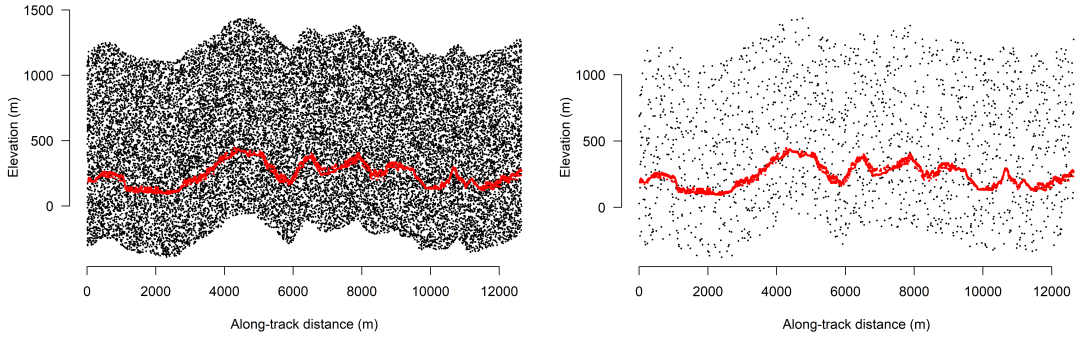
Currently, available data products (Leigh et al. 2015; Popescu et al. 2018) contain numerous amount number of noise photons, especially for the vegetation areas (Brown et al. 2016). Although the classification tags have been provided from the data products, there could still exist some mis-labelled photons for various terrain and atmospheric conditions at a global scale for ATLAS data. Therefore, effective methods to distinguish the correct signal photons from the noise photons are required for further forest application. Previous researches have done some works to distinguish signal photons from these ATLAS-like data, such as the spatial statistical techniques (Herzfeld et al. 2014), an ellipse search area based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (Zhang and Kerekes 2015), the cumulative density-based method (Gwenzi et al. 2016), a particle swarm optimisation-based noise filtering algorithm (Huang et al. 2019), and a ground and top of canopy extraction approach using local outlier factor with ellipse searching area (Chen et al. 2019).

However, these existing methods are mostly based on unsupervised methods, whereas supervised approaches are barely investigated. In general, the purpose of distinguishing the signal photons from noise photons can be considered as a classification problem, and the supervised methods, especially machine learning approaches, have certain advantages over the unsupervised ones: 1) A supervised approach could achieve relatively high accuracy based on limited manually identified training labels, which could be useful for operational processing in a large area. 2) The tuning of parameters of machine learning approaches could be done automatically, while unsupervised methods require manually adjusting the parameters for different study areas. 3) The trained models could be transferable with limited variables from new samples in similar conditions.

Random Forest (RF), an ensemble learning method both for classification and regression, has been proved a good and stable predictor under a variety of cases (Liaw and Wiener 2002; Maxwell, Warner, and Fang 2018). RF runs efficiently on large databases and could handle unbalanced datasets, which suits the unbalanced distribution of the enormous signal and noise photons. Despite the growing number of samples, it typically exhibited good robustness to avoid the over-fitting problem. Therefore, RF is chosen to implement the machine learning based classification for photon counting LiDAR data in this paper.

2. Study Sites and Data

In this paper, the data we used are MATLAS data, which are simulated to fit the expected performance of the ICESat-2 ATLAS instrument by adjusting the existing



(a) (b)

Figure 1. Examples of MATLAS data in West Coast site, the red points stand for potential signal photons identified by the classification flags. (a) High noise level scenario. (b) Low noise level scenario.

Table 1. Description of MATLAS data used in this paper.

Site	Environment	Closure (%)	File name	No. of photons	SNR	Length (m)
East Coast	Temperate hilly	90	t231600_8B_1VEG	45000	0.82	13036
			t231600_8B_2VEG	21943	10.49	13023
			t231900_8B_1VEG	42932	0.80	12390
			t231900_8B_2VEG	20821	10.67	12372
Virginia	Vegetation temperate flat average cover	55	t222500_8a_1VEG	34621	4.70	12922
			t222500_8a_2VEG	54215	1.11	12902
West Coast	Temperate montane	90	t024900_8C_1VEG	43131	0.81	12469
			t024900_8C_2VEG	21037	10.64	12469
			t025000_8C_1VEG	44518	0.80	12887
			t025000_8C_2VEG	21438	10.35	12880
			t025200_8C_1VEG	44208	0.80	12699
			t025200_8C_2VEG	21047	10.34	12699
			t025500_8C_1VEG	43666	0.81	12650
			t025500_8C_2VEG	21232	10.82	12639

MABEL data. To produce MATLAS data, the signal, solar noise, and instrument noise levels are adjusted based on NASA's vegetation design case type first. Next, the spatial variation of signal and noise photons from the original MABEL is preserved. Finally, a large footprint size is formed by combining adjacent channels from the original MABEL data (Hancock 2014).

Table 1 lists the 14 MATLAS datasets from 3 different sites in East Coast, Virginia and West Coast of the USA, representing various vegetation types with different canopy closure fraction and Signal-to-Noise Ratio (SNR). The SNR in each vegetation type is calculated based on the classification flags from the data product themselves provided by NASA. The horizontal coordinates are converted to the along-track distance and the vertical coordinates stand for absolute height. Figure 1 shows two examples of the MATLAS data in West Coast, and it presented numerous noise photons in the atmosphere and even below the ground, especially for the high noise level scenario.

3. Methods

3.1. *Overview*

The machine learning based method we implemented is demonstrated in Figure 2. First, 12 features that could characterise the statistical properties of the photons are

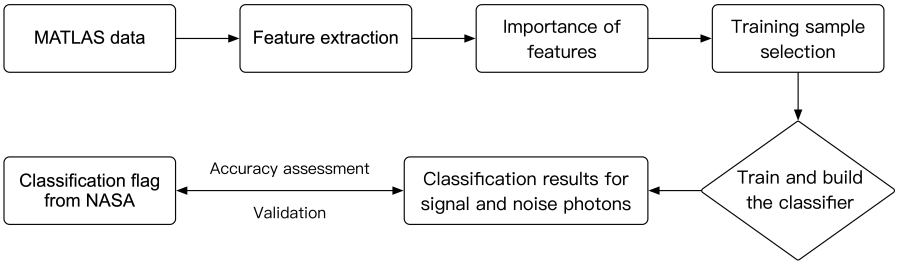


Figure 2. The flowchart of the methods we proposed.

Table 2. Features proposed to train the classifier of photon counting LiDAR data

No.	Feature	Description
1	h	The height of the photon
2	dist	The along-track distance of the photon
3	dist.mean	The difference between height of a photon and the mean value in the surrounding 10 m window
4	dist.median	The difference between height of a photon and the median value in the surrounding 10 m window
5	dist.kmeans	The distance to the corresponding cluster centres of k-means for every photon ($k = 2$)
6	dist.p10	The difference between height of a photon and the 10th percentile in the surrounding 10 m window
7	dist.p25	The difference between height of a photon and the 25th percentile in the surrounding 10 m window
8	dist.p50	The difference between height of a photon and the 50th percentile in the surrounding 10 m window
9	dist.p75	The difference between height of a photon and the 75th percentile in the surrounding 10 m window
10	kNNdist3	The k-nearest neighbours distance for every photon ($k = 3$)
11	h.kurtosis	The difference between kurtosis of a photon and the mean value in the surrounding 10 m window
12	h.skewness	The difference between skewness of a photon and the mean value in the surrounding 10 m window

proposed, after which we rank all 12 variables based on RF modelling for all data in 3 sites to select features with the most representative contributions and stable rankings. Then we determine the numbers of samples used to train the RF classifier to distinguish the signal and noise photons, and further apply the model to the whole coverage of the data. Besides, we investigate the sensitivity of the numbers of samples and the accuracy indicators of the classification results. Finally, we validate the results with the classification flags offered by NASA and evaluate the transferability of the models established.

3.2. *Features extraction and selection*

Table 2 shows the 12 features proposed in our method, including the height and along-track distance, the k-nearest neighbours distance, the distance to the corresponding cluster centres of k-means, the difference between height-related statistical metrics, the mean or percentile values of all the photons at every 10 m window. The mean and percentile features are defined as the following equation:

$$F_i = p_i(x, y) - f[\sum_{i=1}^{n} p_i(x, y)] \tag{1}$$

where F_i represents the feature calculated for the i^{th} photon in every 10 m window size, x and y represent the along-track distance and photon height within the window, and $p_i(x, y)$ stands for the statistical metrics (mean, median, percentiles, kurtosis and skewness) in every 10 m window, respectively. The function $f[\sum_{i=1}^{n} p_i(x, y)]$ is used to

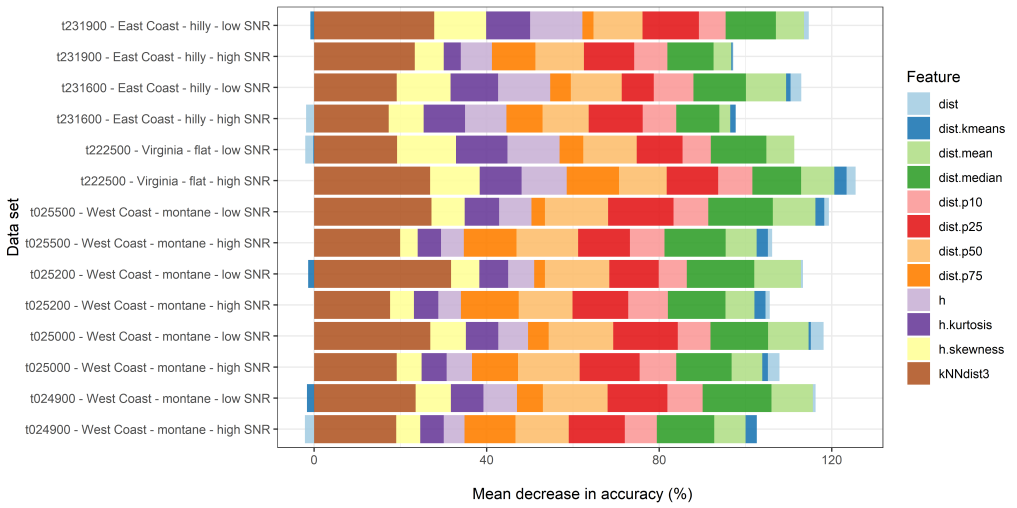


Figure 3. The results of feature importance based on RF model for all the 14 datasets.

calculate the mean or percentile values for the statistical metrics of all photons within the 10 m window.

3.3. *Model establishment and accuracy assessments*

Here, we developed the RF classification model using the features determined by the feature selection results and photons chosen for the training samples. The reference data are the photon classification flags embedded in data products by NASA with a careful visual inspection to correct the obvious noise photons when necessary. Four statistical indicators, namely the accuracy, kappa coefficient, specificity, and F1 score, are calculated from the confusion matrix to examine the results quantitatively. They are defined as follows: accuracy is the proportion of the total number of photons identified as signal and noise photons correctly against the total photon number; kappa coefficient measures the prediction performance of the RF classifier using ground validation; specificity is the proportion of photons considered as noise photons that are correctly identified, and the F1 score is the harmonic mean of precision and sensitivity, in which the precision is the fraction of true signal photons from all points identified as photons and sensitivity is photons considered as signal photons that are correctly identified. Furthermore, the study also investigated the model transferability across data from different sites. We used the dataset in West Coast with a total photon number of 43666 and SNR of 0.81 to train the model and then applied the model to the remaining 13 datasets.

4. Results and Discussion

4.1. *Results of the importance of features*

Figure 3 shows the ranking results of feature importance based on RF for all 14 datasets. It indicated that the contribution of the 12 features varied with different datasets, but some still occupy a large portion in terms of importance. It can be observed that the top 3 features are kNNdist3, dist.median and dist.p50, with the

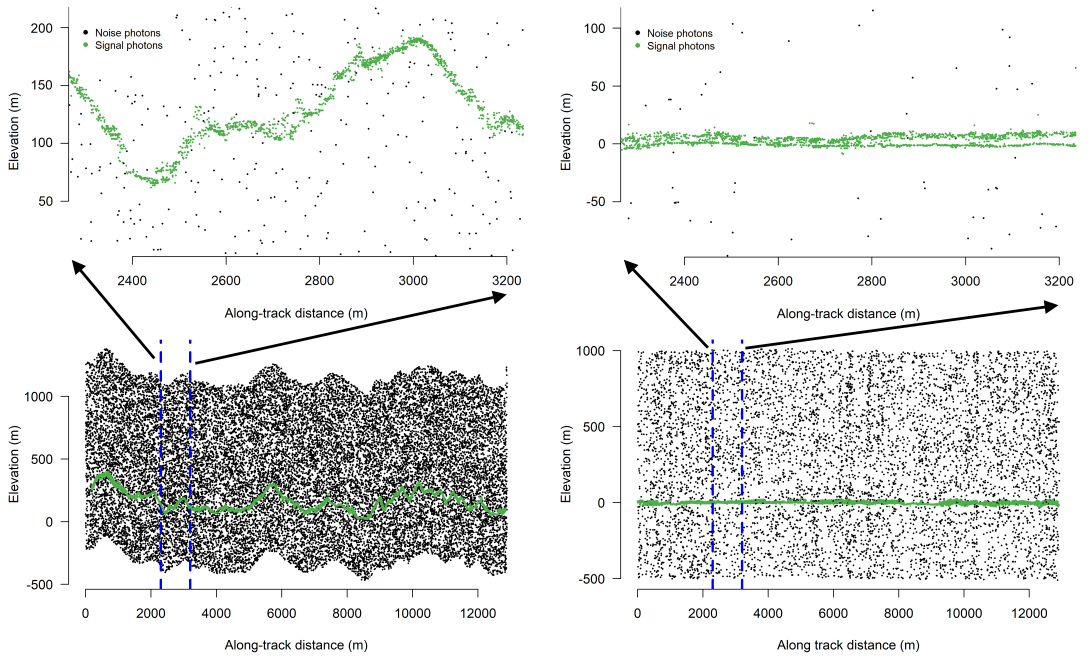


Figure 4. The classification results using RF models. (a) The potential signal photons in West Coast site. (b) The potential signal photons in Virginia site.

accumulated values of mean decrease in accuracies (for all 14 datasets in total) of 319%, 182% and 181% in all datasets, respectively.

It can be seen that the kNN distance has the largest percentage of the model contribution both for every single dataset and the overall results. The dist.median and dist.p50 are the distance features of a photon to the mean value within the 10 m window, indicating the reachability towards the nearby signal centre. In addition, these features present fine stability across different study sites. Consequently, they are selected for further model establishment.

4.2. *Classification results using RF models*

Figure 4 demonstrates the classification results of signal and noise photons using the RF models for datasets of West Coast and Virginia sites, indicating that most of the signal photons have been separated correctly from noise photons using RF classifier. It should also be mentioned that the total number of photons in these two datasets is 44518 and 34621, stretching over a distance of 12887 m and 12922 m, respectively. In the meantime, only 3 features (i.e., kNNdist3, dist.median, and dist.p50) together with 200 photon samples, accounting for 0.45% and 0.58% of photons in the corresponding trajectory, are used here to train the model, and it turns out that small training samples can achieve high modelling accuracy in our study sites.

By visual inspection, it can be implied that the RF model based classifier achieves good results both for data in a low noise rate scenario on flat terrain (Figure 4b) and data in a high noise rate scenario on complex terrain (Figure 4a). Even with a very limited number of samples, the classifier could still distinguish the signal photons from the noise photons very well.

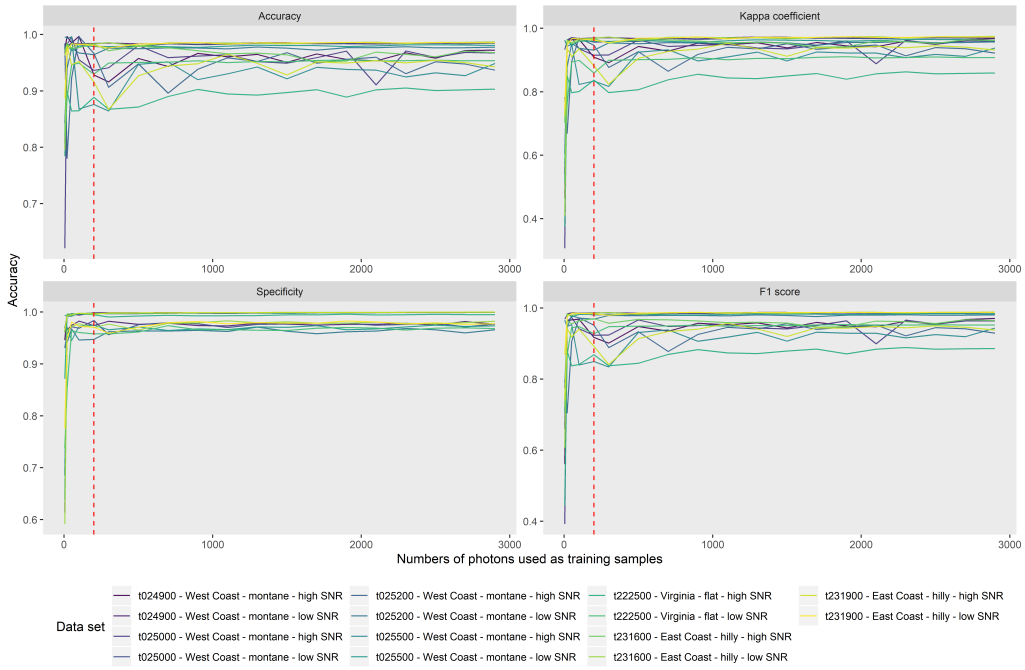


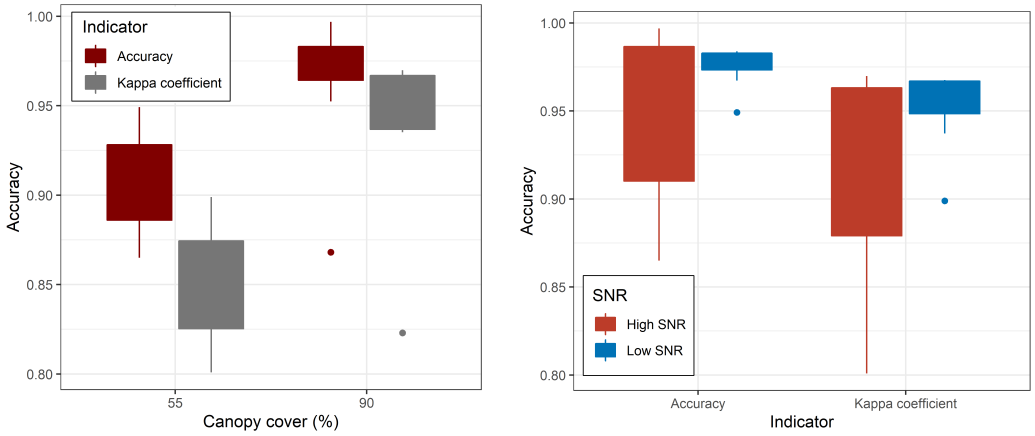
Figure 5. The relationships between the number of samples used to train the model and the four statistical indicators in all 3 study sites. The dashed red line marks 200 photons used as training samples.

4.3. *Accuracy assessment*

The relationships between the number of samples used to train the RF models and the accuracy indicators are shown in Figure 5. The performance of specificity is relatively stable among all four indicators. With samples used to train the RF models increasing from 0 to around 100, the statistical indicators would increase accordingly. However, after the number of samples used is over 200, the four indicators start to fluctuate.

As shown in Figure 5, the accuracy, kappa coefficient, specificity, and F1 score in West Coast site are 0.98, 0.96, 0.98, and 0.98, respectively. Similarly, the corresponding values are 0.89, 0.84, 0.99, and 0.87, respectively for the Virginia site. In terms with all the 14 datasets in 3 sites, the mean values of four indicators are 0.95, 0.92, 0.98, and 0.92 respectively when only 200 samples are fed to the model, which confirms the results of Figure 4 from the quantitative point of view. Notably, 200 samples account for less than 1% of all the photons along the trajectory over around 12 km could achieve high classification performance, indicating the high potential that the model can be readily generalised for large-scale utilisation. Furthermore, we analysed the influence of accuracy for different canopy cover and SNR. Figure 6a shows that a higher canopy cover group has a better overall accuracy, which could be a more substantial portion of vegetation photons increased with a higher closure. Figure 6b shows that the high SNR group has a better overall performance than the low SNR group, which could be a high SNR dataset indicates a relatively low noise scenario.

Overall, our findings suggested that the machine learning approach we implemented is capable of detecting the potential signal photons for MATLAS data, even under varying situations including terrain surfaces, SNR, canopy closure fraction and vegetation type. In addition, our study highlights that only a very small number of samples are required to train the classifier, and this robust model is capable of applying among a relatively large area.



(a) (b)

Figure 6. The influence of different conditions on the accuracy. (a) The influence of canopy cover. (b) The influence of SNR.

4.4. *Model transferability*

Figure 7 shows the transferability of the RF model we developed for the other 13 datasets. Compared with West Coast and East Coast sites, the indicators to evaluate the model transferability are relatively low in Virginia site, of which the lowest one is 0.86, 0.77, 0.98 and 0.83 for accuracy, kappa coefficient, specificity, and F1 score respectively. In contrast, the mean values of the four indicators for West Coast site are 0.96, 0.93, 0.97 and 0.95 respectively, while for East Coast site they are 0.98, 0.94, 0.97 and 0.96 respectively.

The possible explanation of accuracy changes is that the dataset used to build the model is applied for rough terrain in East and West Coast, while the terrain surface in Virginia site is flat. Besides, the SNR in Virginia site is significantly different compared with the other two sites, which might have influences on the model transferability. Therefore, the samples require similar terrain and SNR conditions before being selected for transferring the model. In summary, the RF model has better transferability in West and East Coast site than Virginia site, indicating that there are some limitations to apply the models directly to other sites with significantly different terrain or noise rate conditions.

Through the results obtained by RF classier across different sites, the model transferability is proven to be applicable with satisfactory performance, and it is confirmed that the features selected to train the model are sufficiently independent and representative for more substantial data coverage. These findings suggest that the machine learning based method has the potential for photon counting LiDAR data denoising with the assumption that models are trained under similar data conditions.

5. Conclusion

In this paper, a machine learning based approach was proposed to detect the potential signal photons from the noise photons for photon counting LiDAR data. We found that kNN distance and the reachability towards the nearby signal centre showed high stability and were thereby proved to be suitable features for the model establishment. The relevant quantitative assessment demonstrated that the machine learning approach

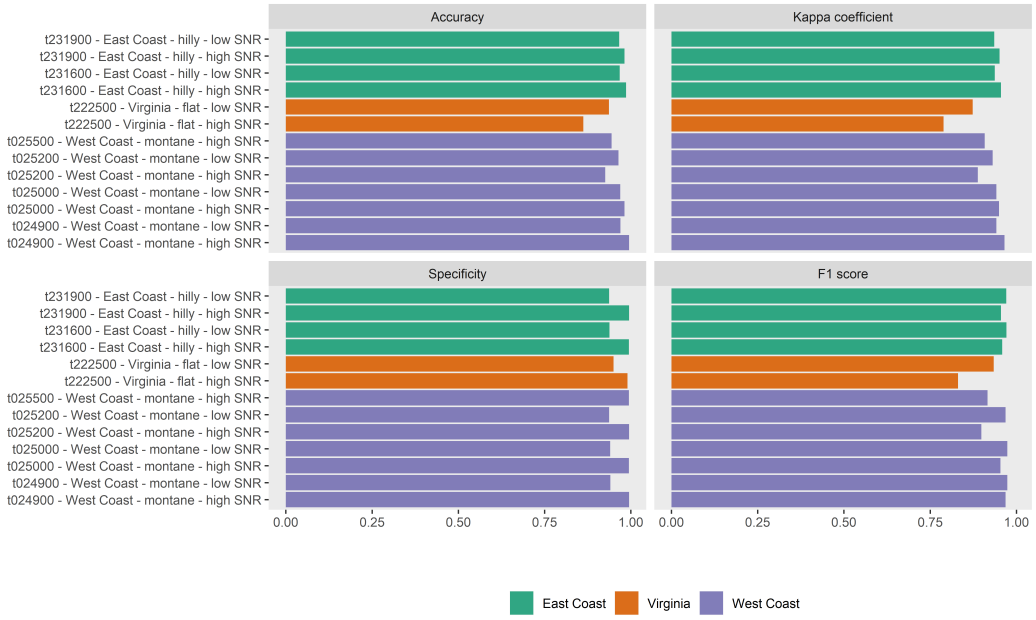


Figure 7. The transferability of the RF model built using the dataset in West Coast sites. The green, orange, and purple bars stand for the accuracies using the developed RF model directly to classify the photons in the East Coast, Virginia, and West Coast sites, respectively.

could distinguish signal photons from the noise photons very well, regarding the high accuracy indicators with a very limited number of samples both in flat or rough terrain conditions. Further analysis proved the potential of model transferability across different sites with similar terrain and SNR conditions. These findings indicated that our methods would be of use for future applications of ICESat-2 vegetation studies.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 41871278 and Grant 31570546. The authors would like to thank the valuable comments and constructive suggestions from the anonymous referees that helped improve the manuscript.

Disclosure statement

The authors declare no conflict of interest.

References

Brown, Molly E, Sabrina Delgodo Arias, Thomas Neumann, Michael F Jasinski, Pamela Posey, Greg Babonis, Nancy F Glenn, Charon M Birkett, Vanessa M Escobar, and Thorsten Markus. 2016. "Applications for ICESat-2 Data: From NASA's Early Adopter Program." *IEEE Geoscience and Remote Sensing Magazine* 4 (4): 24–37.

Chen, Bowei, Yong Pang, Zengyuan Li, Peter North, Jacqueline Rosette, Guoqing Sun, Juan Suárez, Iain Bye, and Hao Lu. 2019. "Potential of Forest Parameter Estimation Using Met-

rics from Photon Counting LiDAR Data in Howland Research Forest." *Remote Sensing* 11 (7): 856.

Duncanson, L I, K O Niemann, and M A Wulder. 2010. "Estimating forest canopy height and terrain relief from GLAS waveform metrics." *Remote Sensing of Environment* 114 (1): 138–154.

Evans, Tyler. 2014. "Optical Development System life cycle for the ICESat-2 ATLAS instrument." In *2014 IEEE Aerospace Conference*, 1–12. IEEE.

Gwenzi, David, Michael A Lefsky, Vijay P Suchdeo, and David J Harding. 2016. "Prospects of the ICESat-2 laser altimetry mission for savanna ecosystem structural studies based on airborne simulation data." *ISPRS Journal of Photogrammetry and Remote Sensing* 118: 68–82.

Hancock, David. 2014. https://icesat-2.gsfc.nasa.gov/icesat2/legacy-data/matlas/docs/.

Herzfeld, Ute Christina, Brian W McDonald, Thomas Allen Neumann, Bruce F Wallin, Thomas A Neumann, Thorsten Markus, Anita Brenner, and Christopher Field. 2014. "Algorithm for detection of ground and canopy cover in micropulse photon-counting lidar altimeter data in preparation for the ICESat-2 mission." *IEEE Transactions on Geoscience and Remote Sensing* 52: 2109 – 2125.

Huang, Jiapeng, Yanqiu Xing, Haotian You, Lei Qin, Jing Tian, and Jianming Ma. 2019. "Particle Swarm Optimization-Based Noise Filtering Algorithm for Photon Cloud Data in Forest Area." *Remote Sensing* 11 (8): 980.

Lefsky, Michael A, David J Harding, Michael Keller, Warren B Cohen, Claudia C Carabajal, Fernando Del Bom Espirito-Santo, Maria O Hunter, and Raimundo de Oliveira. 2005. "Estimates of forest canopy height and aboveground biomass using ICESat." *Geophysical Research Letters* 32 (2): L22S02.

Leigh, Holly W, Lori A Magruder, Claudia C Carabajal, Jack L Saba, and Jan F McGarry. 2015. "Development of Onboard Digital Elevation and Relief Databases for ICESat-2." *IEEE Transactions on Geoscience and Remote Sensing* 53: 2011–2020.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. http://CRAN.R-project.org/doc/Rnews/.

Los, SO, J Rosette, Natascha Kljun, PRJ North, Laura Chasmer, J Suárez, Chris Hopkinson, et al. 2012. "Vegetation height products between 60 S and 60 N from ICESat GLAS data." *Geoscientific Model Development* 5: 413–432.

Markus, Thorsten, Tom Neumann, Anthony Martino, Waleed Abdalati, Kelly Brunt, Beata Csatho, Sinead Farrell, et al. 2017. "The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation." *Remote Sensing of Environment* 190: 260–273.

Maxwell, Aaron E, Timothy A Warner, and Fang Fang. 2018. "Implementation of machine-learning classification in remote sensing: An applied review." *International journal of remote sensing* 39 (9): 2784–2817.

Popescu, SC, T Zhou, R Nelson, A Neuenschwander, R Sheridan, L Narine, and KM Walsh. 2018. "Photon counting LiDAR: An adaptive ground and canopy height retrieval algorithm for ICESat-2 data." *Remote Sensing of Environment* 208: 154–170.

Rosette, Jacqueline, Bruce Cook, Ross Nelson, Chengquan Huang, Jeff Masek, Compton Tucker, Guoqing Sun, et al. 2015. "Sensor Compatibility for Biomass Change Estimation Using Remote Sensing Data Sets: Part of NASA's Carbon Monitoring System Initiative." *IEEE Geoscience and Remote Sensing Letters* 12 (7): 1511–1515.

Yu, Anthony W, Mark A Stephen, Steven X Li, George B Shaw, Antonios Seas, Edward Dowdye, Elisavet Troupaki, Peter Liiva, Demetrios Poulios, and Kathy Mascetti. 2010. "Space laser transmitter development for ICESat-2 mission." In *Proceedings of the SPIE*, Feb., 757809. NASA Goddard Space Flight Ctr., United States.

Zhang, Jiashu, and John Kerekes. 2015. "An Adaptive Density-Based Model for Extracting Surface Returns From Photon-Counting Laser Altimeter Data." *IEEE Geoscience and Remote Sensing Letters* 12: 726–730.