


# Leveraging Airborne LiDAR Data and Gradient Boosting for Mapping the Density of Different Sized Trees

Yuri Shendryk , Member, IEEE, and Emma Gorrod

**Abstract**—Information on the distribution of trees with different diameters at breast height (DBH) is needed to inform management programs aimed at achieving conservation objectives in high stem density forest stands. This article explored the feasibility of mapping the density of trees with different DBHs using airborne LiDAR data. Experiments were conducted in the largest river red gum forest in the world, located in the southeast of Australia. Field measured data on trees with different DBHs were used for the supervised learning of airborne LiDAR scans with a pulse density of 5.92 pulses/m<sup>2</sup>. Specifically, the hyperparameters of gradient boosting and random forest regressors were tuned to produce a viable solution for mapping the density of different sized trees at the plot level. Our results indicate that the total tree density (DBH > 0 cm; height > 1.37 m) can be mapped using airborne LiDAR data with the coefficient of determination  $R^2$  of up to 0.67, with gradient boosting outperforming random forest. However, the accuracy of mapping the density of saplings (DBH ≤ 10 cm), small trees (10 cm < DBH ≤ 50 cm), and large trees (DBH > 50 cm) differed with  $R^2$  of 0.65, 0.60, and 0.42, respectively. These results show that the airborne LiDAR data can provide a viable solution for mapping the density of small trees (DBH ≤ 50 cm) over large areas and has the potential for mapping the density of large trees (DBH > 50 cm).

**Index Terms**—Density, diameter at breast height (DBH), forest, gradient boosting, light detection and ranging (LiDAR), machine learning, random forest.

## I. INTRODUCTION

**F**OREST thickening caused by land management, altered disturbance regimes, and climatic factors is increasingly common globally [1]. In the largest river red gum forest in the world, Barmah–Millewa Forest (BMF) in Australia, stands dominated by the high densities of slender stems with few large trees have become widespread [2]–[5]. After gazettal as

a national park in 2010, multiple management programs were introduced to improve the health of the BMF, such as environmental flows in river systems [6] and an ecological thinning trial [3]. To inform management programs that aim to achieve conservation objectives, a spatial representation of the density of different sized trees overtime is required for BMF. In this respect, the remote sensing technology may provide essential dynamic information on a scale that field-based studies cannot match.

Specifically, airborne light detection and ranging (LiDAR) technology is able to provide vegetation measurements that are highly related to forest attributes (e.g., stem volume, tree density, and above-ground biomass [7]–[10]) at the plot level. Furthermore, recent advances in the LiDAR technology and algorithm development have enabled the estimation of the above-mentioned forest attributes at the individual tree level [11]–[13]. Individual tree detection and segmentation methods generally require LiDAR data with high pulse density (>10 pulses/m<sup>2</sup>). Hence, one of the limitations of the previous plot-level studies utilizing LiDAR data with a relatively low pulse density (<10 pulses/m<sup>2</sup>) was that they did not directly estimate tree diameters at breast height (DBH) [10], [14], [15]. However, the information on the distribution of trees with different DBHs is necessary to assess forest stand properties and estimate growth and future forest prescriptions [16]. Multiple studies achieved high accuracies (the coefficient of determination  $R^2$  of up to 0.86) when predicting the total density or density of large trees in coniferous forests [17], [19]. Relatively low accuracies ( $R^2$  of up to 0.43) were reported when predicting the total tree densities in tropical and eucalypt forests, while no study has investigated the prediction of the density of different sized trees using airborne LiDAR data at the plot level without relying on individual tree detection. Therefore, the aim of this study was to evaluate the ability of airborne LiDAR scans with a pulse density of 5.92 pulses/m<sup>2</sup> to upscale detailed field measurements of different sized (in terms of DBH) trees in the BMF at the plot level. The specific objectives were as follows.

- 1) Explore the feasibility of airborne LiDAR for mapping the density of trees with different DBHs.
- 2) Determine the accuracy with which different tree size classes can be mapped from the air.

To date, the most common methods for predicting tree density at the plot level using airborne LiDAR scans were regression analyses [17]–[19], while machine learning approaches were

Manuscript received July 27, 2020; revised October 8, 2020; accepted December 4, 2020. Date of publication December 21, 2020; date of current version January 8, 2021. This work was supported in part by the NSW Department of Planning, Industry and Environment (DPIE) and in part by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). (Corresponding author: Yuri Shendryk.)

Yuri Shendryk is with Agriculture and Food, Commonwealth Scientific and Industrial Research Organization, Canberra, ACT 2601, Australia (e-mail: yuri.shendryk@gmail.com).

Emma Gorrod is with the Department of Planning, Industry and Environment, Department of Planning Industry and Environment, Parramatta, NSW 2124, Australia (e-mail: emma.gorrod@environment.nsw.gov.au).

Digital Object Identifier 10.1109/JSTARS.2020.3046303

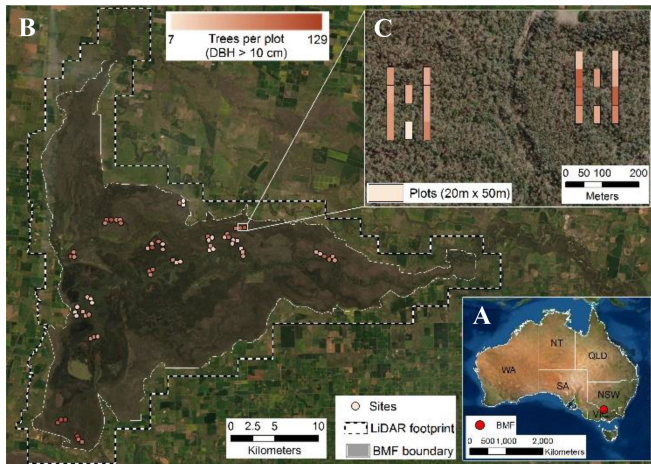


Fig. 1. Study area. (A) Location of the BMF. (B) Distribution of field sites and LiDAR footprints across BMF. (C) Example of two sites containing 20 field plots (20 × 50 m).

less commonly used [10], [16]. Therefore, in this study we evaluated two popular decision tree-based machine learning algorithms, gradient boosting [20] and random forest [21], which require minimum data preprocessing and are able to take advantage of the data dimensionality. The decision tree-based machine learning algorithms have shown good potential in multiple classification benchmarks [22], [23] and, while being more interpretable, consistently outperform neural network models on tabular-style datasets, where features are individually meaningful [24].

## II. METHODOLOGY

### A. Study Area

The BMF [see Fig. 1(A)] occupies 737 km<sup>2</sup> and is the largest contiguous area of *Eucalyptus camaldulensis*, commonly known as river red gums, in the world. The BMF is a floodplain forest that spreads along the Murray river, with Millewa forest on the northern side of the Murray river in the state of New South Wales, and Barmah forest on the southern side of the river in the state of Victoria [25]. This forest complex consists of river red gum forest (71%), river red gum woodland (23%), and mixed box eucalypt woodland (6%) [26]. The BMF is structurally complex with highly variable age and health conditions as well as tree densities ranging from more than 4000 trees/ha in forests to less than 50 trees/ha in open woodlands [27], [28].

### B. LiDAR Data

Airborne LiDAR scans covering the whole extent of the BMF were downloaded from the Geoscience Australia elevation information system (ELVIS) [29]. The LiDAR scans were acquired between 10 September and 7 November 2015 using Trimble AX60 system in a full-waveform mode with parameters specified in Table I. Ground control was also collected between 25 September and 6 October 2015 and used to verify the accuracy of LiDAR scans. Comparing LiDAR scans with the ground control points resulted in a calculated accuracy of 0.166 m at two sigmas

TABLE I  
AIRBORNE LIDAR ACQUISITION PARAMETERS

Sensor name	Trimble AX60
Scanning mechanism	Rotating polygon mirror
Wavelength	1062 nm
Scan angle range	±30°
Laser pulse repetition rate	400 kHz
Beam divergence	≤0.25 mrad
Flying height above ground	850 m
Swath width	981 m
Swath overlap	30%
Footprint diameter	22 cm
Pulse <sup>1</sup> density	5.92 pulses/m <sup>2</sup>
Point <sup>2</sup> density	9.95 points/m <sup>2</sup>
Pulse spacing	0.41 m
Point spacing	0.32 m
Spatial accuracy (horizontal / vertical)	0.8 m / 0.3 m
Return count	up to 7
Horizontal / vertical datum	GDA94 / AHD71

<sup>1</sup>Pulse is the laser signal sent out from the LiDAR system toward the ground.

<sup>2</sup>Point is the signal or multiple signals reflected from target(s) back toward the LiDAR system.

(i.e., two standard deviations). The inertial measurement unit (IMU) (Trimble AP50 GNSS/IMU) and postprocessed airborne GPS logs were used to generate the LiDAR point cloud from the waveform instrument data [29].

The LiDAR data used in this study were originally acquired to provide a detailed terrain surface for modeling the flows and volumes of water within the floodplains and channels of the Edward–Wakool region [30]. An automatic classification algorithm applied in TerraScan software to produce a classification of LiDAR scans by the vendor for Vaze *et al.* [30] study was deemed inappropriate for this study due to excessive occurrence of unclassified low (below ground surface) and high (above vegetation) noise points. LiDAR scans were composed of multiple flight lines that had a nominal swath overlap of 30%. However, there were also areas that were surveyed twice (potentially due to scanning gaps) leading to the swath overlaps of 100% in some areas. Major issues in overlapping areas of LiDAR swaths were noted in [30] (e.g., the elevation differences of over ±30 cm) and were also confirmed in this study, suggesting that flight lines were not properly aligned by the vendor. The misalignment of swaths could lead to inaccurate point classification as well as “stripping” issues when calculating the forest density metrics in areas of swaths overlap. Therefore, in this study, LiDAR scans were reprocessed and reclassified using LASTools software [31].

First, LASTools software was used to recover the original 123 flight lines and to minimize swath overlaps by preserving the points within 0.3 m spacing with the lowest absolute scan angle. This led to an average point density decrease from 9.95 to 7.23 points/m<sup>2</sup>, and pulse density decrease from 5.92 to 4.45 pulses/m<sup>2</sup>. This also led to an average point spacing increase from 0.32 to 0.37 m, and pulse spacing increase from 0.41 to 0.47 m. Reprocessed flight lines were then merged and split into 2 × 2 km tiles with a 50 m buffer to speed up further processing and avoid classification artifacts on the edges of each tile. Duplicate, low, and high noise points were automatically

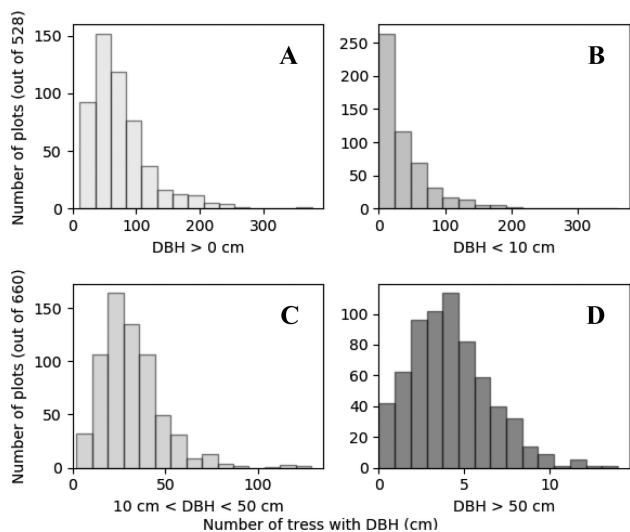


Fig. 2. Distribution of tree density with different DBHs (i.e., response variables). (A) “all trees” (DBH > 0 cm). (B) “saplings” (DBH ≤ 10 cm). (C) “small trees” (10 cm < DBH ≤ 50 cm). (D) “large trees” (DBH > 50 cm).

classified and removed from further processing. An automatic classification algorithm [31] was further applied to produce a classification of ground points, which were used to height normalize the remaining unclassified points. No position and orientation files were available for LiDAR data in the ELVIS portal [29], which prevented the performance of LiDAR swaths alignment. Similarly, no range information was available in LiDAR data preventing the correction of intensity values.

### C. Field Data and Response Variables

The field measurements of the density of different sized trees were collected in June and July 2016 before a major flooding event and were completed in February 2017 once flood waters had subsided. Field data were collected at 66 sites [see Fig. 1(B)], each site consisting of ten 20 × 50 m plots [Fig. 1(C)] delineated using the standard GPS unit with the positional accuracy of approx. 5 m. The data for each plot consisted of the number of trees (≥ 1.37 m in height) in 10 cm DBH size classes. For the purpose of this study, the numbers of trees were binned into four tree densities with different DBH classes: “all trees” (DBH > 0 cm); “saplings” (DBH ≤ 10 cm); “small trees” (10 cm < DBH ≤ 50 cm); “large trees” (DBH > 50 cm), which are further referred to as response variables. While all of the 660 plots were within the LiDAR footprint [see Fig. 1(B)], only 528 plots had information on “all trees” and “saplings” classes, as in some plots, the information on the number of trees with DBH ≤ 10 cm was not recorded. It is also important to note that 100 out of 660 plots were surveyed for stand structure after they were thinned in 2017. The tree density estimates for these 100 plots were based on stump counts and a tapering equation to allocate each tree to an appropriate size class.

The distributions of all response variables were right skewed (see Fig. 2), which posed a class imbalance problem for machine learning regressors. Therefore, prior to the machine learning

TABLE II  
TOTAL OF 78 PREDICTOR VARIABLES EXTRACTED FROM THE LiDAR SCANS

Predictors	Description
<i>max</i>	maximum height above 0.5 m
<i>avg</i>	average height above 0.5 m
<i>qav</i>	average square height <sup>1</sup> above 0.5 m,
<i>std</i>	standard deviation of height above 0.5 m
<i>ske</i>	height skewness above 0.5 m
<i>kur</i>	height kurtosis above 0.5 m
<i>cov</i>	canopy cover - the number of first returns above 0.5 m divided by the number of all first returns
<i>dns</i>	canopy density - the number of all points above 0.5 m divided by the number of all returns
<i>hom</i>	height of median energy <sup>2</sup>
<i>p05 to p95</i>	5 <sup>th</sup> to 95 <sup>th</sup> percentile height value of all points between 0.5 m and the maximum height
<i>b05 to b95</i>	5 <sup>th</sup> to 95 <sup>th</sup> bincile <sup>3</sup> above 0.5 m
<i>d00 to d21</i>	relative height density - the ratio of total number of points in 2 m intervals (from 0 to 46 m <sup>4</sup> ) divided by the total number of points above 0.5 m
<i>vc00 to vc06</i>	Vertical complexity index <sup>5</sup> for six vertical bin sizes: 0.5 m, 1 m, 2 m, 3 m, 4 m and 5 m starting 0.5 m above the ground
<i>int_max</i>	maximum intensity above 0.5 m
<i>int_avg</i>	average intensity above 0.5 m
<i>int_std</i>	standard deviation of intensity above 0.5 m
<i>int_ske</i>	intensity skewness above 0.5 m
<i>int_kur</i>	intensity kurtosis above 0.5 m

<sup>1</sup>Average square height is an arithmetic mean of squared height. It is a measure of central tendency, and in forestry research, it is considered to be more appropriate than the arithmetic mean height [33].

<sup>2</sup>Height of median energy [7] was calculated by ordering all points above 0.5 m by their elevation. Then, the height was computed at which the sum of intensities of points below and the sum of intensities of points above was identical.

<sup>3</sup>Fraction of return points between the *n*th (i.e., 5th–95th) percentile height and the maximum height (%).

<sup>4</sup>Tallest tree within 20 × 50 m plots was 44.8 m.

<sup>5</sup>Vertical complexity index provides information about the vertical distribution of the points [32].

analysis, all response variables were log transformed to the normal distribution.

### D. Predictor Variables

LAStools software was further used to generate 78 forestry metrics (i.e., predictor variables) within 20 × 50 m plots from height-normalized LiDAR scans (see Table II), some of which were previously shown to be integral for predicting tree biomass, diameter, and basal area at the plot level [7], [16], [18], [32].

Only 6 out of 78 predictors (i.e., *hom*, *int\_max*, *int\_avg*, *int\_std*, *int\_ske*, and *int\_kur*) were calculated using the LiDAR intensity information, as there was pronounced “stripping” noise in intensity data in the overlapping areas of LiDAR swaths.

To investigate the effect of the positional accuracy of delineated plots on the prediction of tree densities, each 20 × 50 m plot was shifted in four cardinal directions (i.e., north, west, east, and south) (see Fig. 3) from its center by 5 m (i.e., approx. accuracy of a GPS receiver used for delineating plots). Each resulting plot area was used to compute 78 predictor variables (see Table II) from height-normalized LiDAR point clouds. Finally, the average of each predictor variable across five plot areas (i.e., center, north, west, east, and south) was

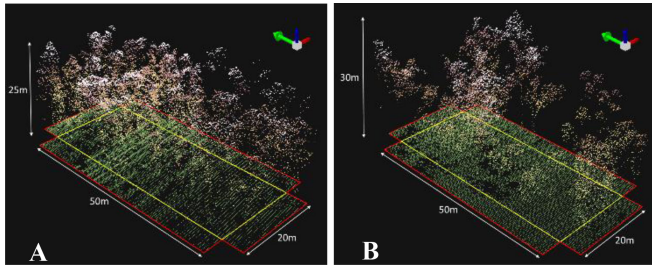


Fig. 3. Examples of LiDAR scans extracted within  $20 \times 50$  plots. (A) Plot containing 112 trees with  $DBH > 10$  cm. (B) Plot containing nine trees with  $DBH > 10$  cm. The central plot is highlighted in yellow, while the plot shifts of 5 m in four cardinal directions are highlighted in red.

TABLE III  
HYPERPARAMETER RANGES USED FOR OPTIMIZING GRADIENT BOOSTING AND RANDOM FOREST REGRESSORS (SEE SCIKIT-LEARN LIBRARY DOCUMENTATION [34] FOR HYPERPARAMETER DESCRIPTION)

Hyper-parameter	Gradient boosting	Random forest
$n\_estimators$	500, 1000, 1500	500, 1000, 1500
$loss$	"ls", "lad", "huber", "quantile"	N/A
$criterion$	'friedman_mse', 'mse'	'mse', 'mae'
$learning\_rate$	0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0	N/A
$max\_depth$	1 – 10 (step: 1)	1 – 10 (step: 1)
$min\_samples\_split$	2 – 20 (step: 1)	2 – 20 (step: 1)
$min\_samples\_leaf$	1 – 20 (step: 1)	1 – 20 (step: 1)
$max\_features$	0.05 – 1 (step: 0.05)	0.05 – 1 (step: 0.05)
$subsampling$	0.5 – 1 (step: 0.05)	N/A
$alpha$	0.75, 0.8, 0.85, 0.9, 0.95, 0.99	N/A

also calculated to investigate whether the averaging of predictor variables could lead to improved prediction accuracy.

### E. Machine Learning

In this study, both gradient boosting [20] and random forest [21] regressors were used to estimate the density of different sized trees. Gradient boosting is an ensemble learning method that combines the predictive power of multiple decision trees using a boosting algorithm. In boosting, the decision tree that improves the model most is added to an ensemble at each iteration until the set number of estimators (i.e.,  $n\_estimators$ ) has been achieved. In contrast to bagging techniques, such as random forest, in which the trees are grown to their maximum extent, boosting makes use of shallow trees with fewer splits and relies on multiple hyperparameters (see Table III) to control the learning process.

To train gradient boosting and random forest regressors with the highest predictive accuracy of tree densities with different DBHs, a Scikit-learn library was used. Scikit-learn is a free software machine learning library for the Python programming language featuring various classification, regression, and clustering algorithms [34]. For each response variable, the data were split into training (80%) and test (20%) sets. Using the training dataset for each response variable, the best value of each hyperparameter (see Table III) was determined using 10000

iterations of randomly generated gradient boosting and random forest pipelines. Each iteration used fivefold cross-validation stratified according to the values of a response variable. At each fold of the cross-validation procedure, 80% of the training dataset was used to train the model and 20% to validate it. The best model according to the highest average cross-validation score was further evaluated using the test dataset, while the final machine learning models for inference were trained on all data.

The accuracy of gradient boosting and random forest models for predicting response variables was assessed using the coefficient of determination  $R^2$  at both training and test stages, while the root-mean-square error (RMSE) was additionally calculated to evaluate final models on the test data. RMSE shows how much predictions deviate from the actual values in the dataset on average and was calculated as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{x}_i - x_i)^2}$$

where  $x_i$  is the measured value,  $\hat{x}_i$  is the predicted value, and  $N$  is the number of measurements.

Through a randomized search on hyperparameters, a total of 48 final machine learning models were selected for different sized tree density prediction [i.e., two machine learning algorithms  $\times$  four response variables  $\times$  six plot areas (i.e., center, north, west, east, south, and average) used for predictor variable extraction].

While the training and optimization of machine learning models were based on predictor variables that were calculated within  $20 \times 50$  m ( $1000$  m<sup>2</sup>) plots, the inference was based on predictor variables calculated within  $32 \times 32$  m grid cells ( $1024$  m<sup>2</sup>) across the entire BMF extent to approximate the size of field plots.

### F. Predictor Importance

Decision tree-based machine learning models have been extensively used in remote sensing to make predictions based on the sets of input predictors. For these applications, models often must be both accurate and interpretable, where interpretability means that we can understand how the model uses input predictors to make predictions. However, widely used explanation methods for decision tree-based models are inconsistent. For example, commonly used information gain [35] is biased as it, first, averages the contribution of predictors across all instances they appear in the trees. This dilutes the calculated importance of some predictors, which are used as a splitter many times, although not always improving the model by a large amount each time it is used, and second, alters the impact of predictors based on their tree depth (i.e., the information gain method is biased to attribute more importance to lower splits) [36]. Therefore, in this study, a recently introduced SHapley Additive exPlanations (SHAP) [36] method to explain the output of the machine learning models was used. SHAP is a game-theoretic approach to explain the output of any machine learning model using Shapley values [37]. The Shapley value is the average marginal contribution of a predictor across all possible coalitions, which

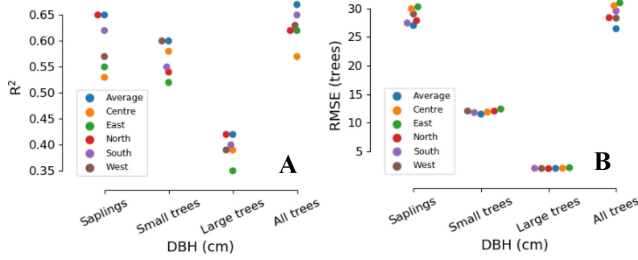


Fig. 4. Results of predicting the density of different sized trees using gradient boosting. The variations in terms of (A)  $R^2$  and (B) RMSE depending on six plot areas (i.e., average, center, east, north, south, and west) used for the predictor variables extraction.

indicates the magnitude of the predictor variable’s contribution to a response variable.

### III. RESULTS

Overall, gradient boosting outperformed random forest in predicting the densities of different sized trees when using predictor variables averaged across five plot areas (i.e., center, north, west, east, and south). Random forest was able to predict the densities of “saplings,” “small trees,” “large trees,” and “all trees” with  $R^2$  (RMSE) of 0.65 (29.45), 0.58 (11.63), 0.39 (2.07), and 0.65 (28.37), respectively. In contrast, gradient boosting predicted the densities of “saplings,” “small trees,” “large trees,” and “all trees” with  $R^2$  (RMSE) of 0.65 (27.00), 0.60 (11.52), 0.42 (2.06), and 0.67 (26.43), respectively. Given a slight superiority of the latter method, the rest of this section will be presented with the results from the gradient boosting analysis.

The accuracy ( $R^2$ ) of “saplings,” “small trees,” “large trees,” and “all trees” predictions using gradient boosting and considering the positional shifts of the plots in four cardinal directions ranged between 0.53 and 0.65, 0.52 and 0.60, 0.35 and 0.42, and 0.57 and 0.67, respectively (see Fig. 4).

The positional shifts of the plots in four cardinal directions generally resulted in  $R^2$  decrease of up to 0.12 (e.g., “saplings”). However, the most accurate models were built when using predictor variables averaged across five plot areas (i.e., center, north, west, east, and south). Fig. 5 shows the predictive performance of final gradient boosting models on a test set (i.e., hold-out sample of 20%) trained using the predictor variables averaged across five plot areas. The gradient boosting models overpredicted the large values and underpredicted the low values of tree densities (see Fig. 5).

The most important predictors of each response variable (as determined by SHAP values) commonly included forest density metrics (e.g., *cov* and *dns*) (see Fig. 6). Moreover, *d01*, *d05*, and *d11* predictors were one of the most important in predicting the densities of “saplings,” “small trees,” and “large trees,” respectively. The high values of *cov* predictor tended to predict high tree densities for every response variable, while high *p05* values increased the chance of predicting the low densities of “all trees” and “saplings” [as indicated by dot color in Fig. 6(A) and (B)]. Interestingly, high *int\_avg* values increased the chance of predicting the low densities of “small trees” [see Fig. 6(C)],

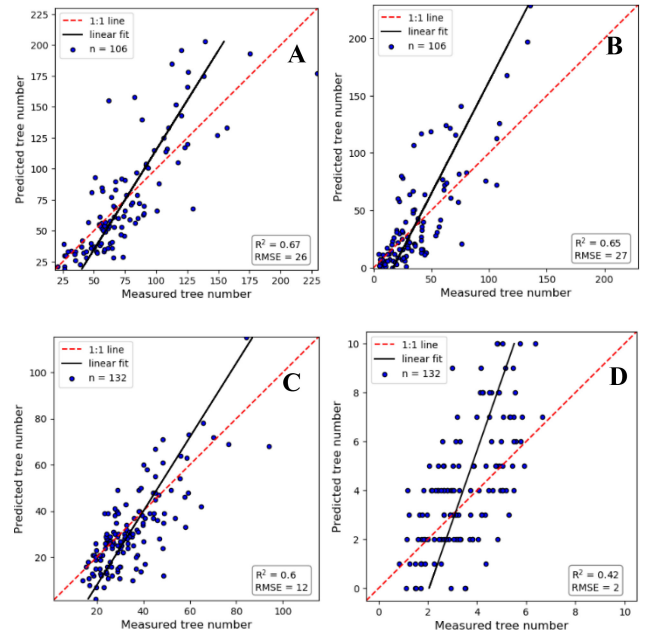


Fig. 5. Accuracy ( $R^2$ ) of the predicted tree density of: (A) “all trees” (DBH > 0 cm), (B) “saplings” (DBH < 10 cm), (C) “small trees” (10 cm < DBH < 50 cm), and (D) “large trees” (DBH > 50 cm) on a test set (i.e., hold-out sample of 20%) using the final gradient boosting models optimized through hyperparameter tuning.

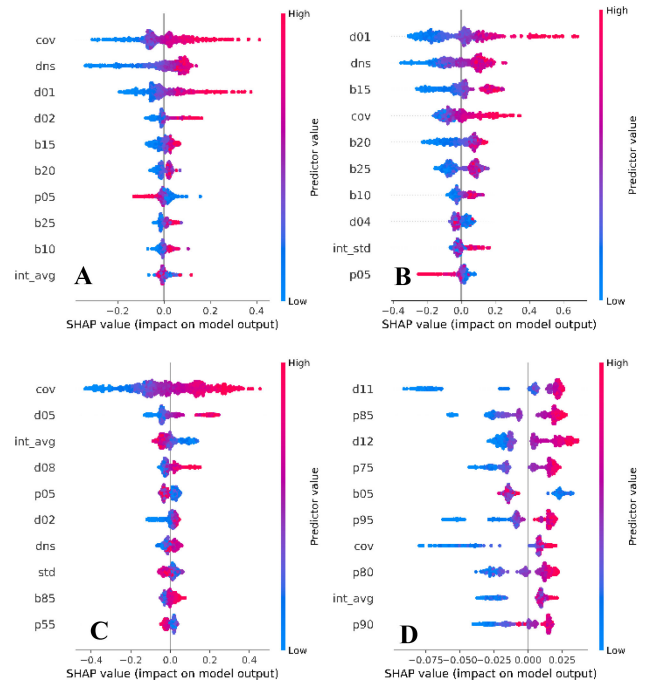


Fig. 6. Predictor importance for the predicted tree densities of (A) “all trees” (DBH > 0 cm), (B) “saplings” (DBH < 10 cm), (C) “small trees” (10 cm < DBH < 50 cm), and (D) “large trees” (DBH > 50 cm). Every field plot has one dot on each row. The x-axis position of the dot is the impact of that predictor on the model’s prediction for the plot, and the color of the dot represents the value of that predictor for the plot. The dots that do not fit on the row pile up to show density. All SHAP values have the same unit, i.e., the unit of the prediction space, and the sum of SHAP values yields the difference of actual and average prediction. The 10 (out of 78) most important predictors for each response variable are shown.

while it had the opposite effect on the predictions of the tree densities of “large trees” [see Fig. 6(D)].

The prediction maps of the density of different sized trees at  $32 \times 32$  m resolution for BMF areas with trees higher than 2 m (i.e.,  $\max > 2$  m) and relatively low vegetation density under 2 m (i.e.,  $d00 < 0.2$ ) ( $712 \text{ km}^2$ , which is 97% of the total BMF area) are presented in Fig. 7. The areas (i.e.,  $32 \times 32$  m plots) within BMF with  $\max < 2$  m and  $d00 > 0.2$  were masked out to exclude swampy areas with very tall phragmites and juncus. Generally, high tree densities were concentrated along the water bodies in the BMF.

#### IV. DISCUSSION

The results of this study indicate that LiDAR-derived predictors were able to estimate the density of trees with  $\text{DBH} > 0$  cm at the plot level with  $R^2$  of up to 0.67. This is a substantial improvement in comparison with the previous research, suggesting that the tree density ( $\text{DBH} \geq 10$  cm) in BMF could be predicted through the random forest regressor with  $R^2$  of 0.48–0.52 when using the airborne low pulse density (approx.  $0.5 \text{ pulses/m}^2$ ) LiDAR data and  $R^2$  of 0.61 when using satellite multispectral and synthetic-aperture radar (SAR) imagery in combination with the airborne low pulse density LiDAR data [38]. However, the improvement in  $R^2$  in this study was likely attributed to the increase in the pulse density (from 0.5 to  $5.92 \text{ pulses/m}^2$ ) rather than the use of gradient boosting over the random forest. It was found that gradient boosting consistently outperformed random forest ( $R^2$  increase of up to 0.03) in predicting tree densities across a range of DBHs. Previously, it was also reported that the accuracy of tree density prediction using LiDAR data was highly dependent on the changes in the pulse density up to  $2 \text{ pulses/m}^2$  and not the choice of a machine learning algorithm [10].

The accuracies of predicting tree densities in this study were difficult to compare with the previous studies relying on airborne LiDAR data. For example, the plot-level density of trees with  $\text{DBH} > 100$  cm was previously predicted with  $R^2 = 0.79$  [17], while the density of all trees was estimated with  $R^2 = 0.53$  [17],  $R^2 = 0.86$  [19], and  $R^2 = 0.79$  [18]. All above-mentioned studies [17]–[19] used multiple linear regression analysis for their predictions; however, they focused on coniferous and mixed-wood forests and failed to report the pulse density of their LiDAR data making any comparison to our study inadequate. In contrast, much lower accuracies were reported when predicting the plot-level tree density in tropical ( $R^2 = 0.43$  [14]) and eucalypt forests ( $R^2 = 0.41$  [15]) using LiDAR data with the point density of up to  $3.5 \text{ points/m}^2$ .

General overprediction of large values and underprediction of low values of tree densities (see Fig. 5) are surprising results, given that the opposite trend is intrinsic to regression tree-based machine learning models [39]. This could be attributed to the right-skewed nature of the tree size class data (see Fig. 2). Postprocessing by adjusting the slope and intercept of the machine learning model output could be used to reduce this bias [39]. Alternatively, a different accuracy metric (e.g., RMSE) and custom loss function that penalizes more heavily overpredictions

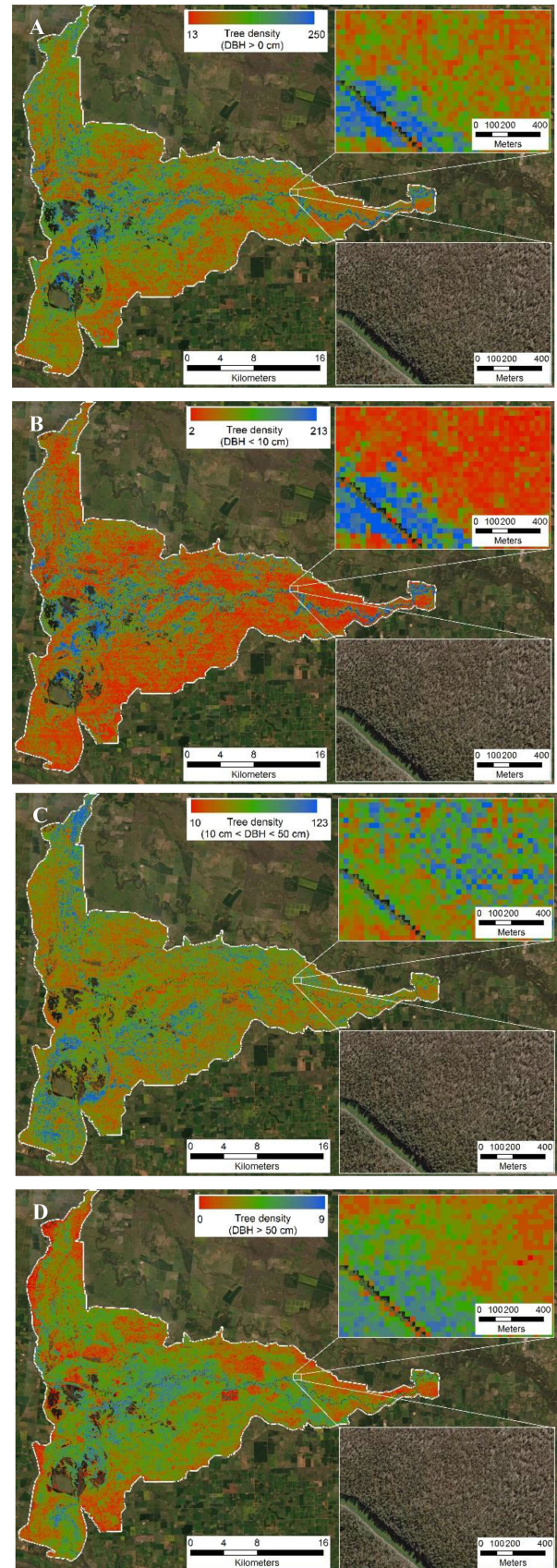


Fig. 7. LiDAR-derived distribution of the density of (A) “all trees” ( $\text{DBH} > 0$  cm), (B) “saplings” ( $\text{DBH} < 10$  cm), (C) “small trees” ( $10 \text{ cm} < \text{DBH} < 50$  cm), and (D) “large trees” ( $\text{DBH} > 50$  cm).

of large values and underprediction of small values could be employed during the model training.

The accuracy of final models for predicting the density of “large trees” (i.e.,  $DBH > 50$  cm) was lower ( $R^2 < 0.42$ ) than that of “small trees” (i.e.,  $DBH \leq 50$  cm and  $R^2 < 0.65$ ), suggesting that the gradient boosting regressor had difficulty teasing out the “signature” of large trees from a LiDAR-derived point cloud with the pulse density of 5.92 pulses/m<sup>2</sup>.

The most important predictors were associated with the forest density metrics (e.g., *dns*, *cov*, and *d00–d21*), which were most affected by “stripping” issues in the areas of LiDAR swaths overlap. Furthermore, according to the importance of predictors associated with forest density (*d00–d21*), the most critical regions of the LiDAR point cloud for predicting the density of “saplings” ( $DBH \leq 10$  cm), “small trees” (i.e.,  $DBH \leq 50$  cm), and “large trees” (i.e.,  $DBH > 50$  cm) were 2–4 m (*d01*), 10–12 m (*d05*), and 22–24 m (*d11*), respectively. Although the intensity-derived predictors (e.g., *int\_avg*) were uncalibrated in this study, they showed to be among the most important in predicting the density of different sized trees.

Overall, the results of this study could be improved after remeasuring the position of existing 660 field plots using a differential GPS system with a  $< 0.1$  m positional accuracy, as the low positional accuracy ( $\sim 5$  m) of field plots substantially affected the accuracy of predicted densities of different sized trees (see Fig. 4). While the positional shifts of the plots in four cardinal directions generally resulted in  $R^2$  decrease of up to 0.12 when predicting tree densities, the highest accuracies were achieved when using predictor variables averaged across five plot areas (i.e., center, north, west, east, and south). Furthermore, the retrieval of orientation and position information associated with the LiDAR data will allow the alignment of swaths, thus reducing elevation differences in the areas of LiDAR swaths overlaps. The availability of orientation and position information would also allow the calibration of intensity noise that primarily appeared in the overlapping LiDAR swaths. Finally, the introduction of additional predictor variables extracted from, for example, other readily available remote sensing datasets for the BMF (e.g., eight-band multispectral WorldView-2 imagery at 0.5 m spatial resolution and ALOS PALSAR SAR imagery at 10 m spatial resolution [38]) could improve the accuracy of results achieved in this study.

## V. CONCLUSION

The ability to map the density of different sized trees is necessary for effective forest management. Our results showed that LiDAR scans can provide a viable solution for mapping the density of “saplings” ( $DBH < 10$  cm and  $R^2 < 0.65$ ) and “small trees” ( $DBH < 50$  cm and  $R^2 < 0.60$ ) over large areas and has the potential for mapping the density of “large trees” ( $DBH > 50$  cm and  $R^2 < 0.42$ ). Given airborne LiDAR scans with the pulse densities of up to 10 pulses/m<sup>2</sup> are available for most of the floodplain eucalypt forests in Australia, a national, wall-to-wall mapping of tree densities could be achieved using the machine learning models developed in this study.

## ACKNOWLEDGMENT

The authors would like to thank the New South Wales Department of Planning, Industry, and Environment (DPIE) for commissioning this research in the context of forest health decline in the Barmah–Millewa Forest, and also D. McAllister and E. Curtis (both of DPIE) for supervising the collection of field data.

## REFERENCES

- [1] A. Rautiainen, I. Wernick, P. E. Waggoner, J. H. Ausubel, and P. E. Kauppi, “A national and international analysis of changing forest density,” *PLoS One*, vol. 6, no. 5, May 2011, Art. no. e19577.
- [2] S. C. Cunningham *et al.*, “A robust technique for mapping vegetation condition across a major river system,” *Ecosystems*, vol. 12, no. 2, pp. 207–219, 2009.
- [3] “Ecological thinning trial in NSW and Victorian river red gum reserves (Experimental design and monitoring plan),” Office Environ. Heritage, Sydney, NSW, Australia, 2012.
- [4] R. M. Nally, S. C. Cunningham, P. J. Baker, G. J. Horner, and J. R. Thomson, “Dynamics of Murray–Darling floodplain forests under multiple stressors: The past, present, and future of an Australian icon,” *Water Resour. Res.*, vol. 47, no. 12, 2011, Art. no. W00G05.
- [5] E. J. Gorrod *et al.*, “Can ecological thinning deliver conservation outcomes in high-density river red gum forests? Establishing an adaptive management experiment,” *Pac. Conserv. Biol.*, vol. 23, no. 3, pp. 262–276, 2017.
- [6] “Barmah–Millewa forest: Environmental water management plan,” Murray–Darling Basin Authority, Canberra, ACT, Australia, 2012.
- [7] J. B. Drake, R. O. Dubayah, R. G. Knox, D. B. Clark, and J. B. Blair, “Sensitivity of large-footprint lidar to canopy structure and biomass in a neotropical rainforest,” *Remote Sens. Environ.*, vol. 81, no. 2/3, pp. 378–392, Aug. 2002.
- [8] M. Dalponte, N. C. Coops, L. Bruzzone, and D. Gianelle, “Analysis on the use of multiple returns LiDAR data for the estimation of tree stems volume,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 2, no. 4, pp. 310–318, Dec. 2009.
- [9] J. J. Richardson and L. M. Moskal, “Strengths and limitations of assessing forest density and spatial configuration with aerial LiDAR,” *Remote Sens. Environ.*, vol. 115, no. 10, pp. 2640–2651, Oct. 17, 2011.
- [10] M. K. Jakubowski, Q. Guo, and M. Kelly, “Tradeoffs between lidar pulse density and forest measurement accuracy,” *Remote Sens. Environ.*, vol. 130, pp. 245–253, Mar. 15, 2013.
- [11] W. Yao, P. Krzystek, and M. Heurich, “Tree species classification and estimation of stem volume and DBH based on single tree extraction by exploiting airborne full-waveform LiDAR data,” *Remote Sens. Environ.*, vol. 123, pp. 368–380, Aug. 2012.
- [12] M. Dalponte, L. Bruzzone, and D. Gianelle, “A system for the estimation of single-tree stem diameter and volume using multireturn LIDAR data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 7, pp. 2479–2490, Jul. 2011.
- [13] T. Allouis, S. Durrieu, C. Végé, and P. Couteron, “Stem volume and above-ground biomass estimation of individual pine trees from LiDAR data: Contribution of full-waveform signals,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 924–934, Apr. 2012.
- [14] M. W. Palace *et al.*, “Estimating forest structure in a tropical forest using field measurements, a synthetic model and discrete return lidar data,” *Remote Sens. Environ.*, vol. 161, pp. 1–11, May 2015.
- [15] A. Haywood and C. Stone, “Using airborne laser scanning data to estimate structural attributes of natural eucalypt regrowth forests,” *Aust. Forestry*, vol. 74, no. 1, pp. 4–12, Mar. 2011.
- [16] J. L. Strunk, P. J. Gould, P. Packalen, K. P. Poudel, H.-E. Andersen, and H. Temesgen, “An examination of diameter density prediction with k-NN and airborne lidar,” *Forests*, vol. 8, no. 11, 2017, Art. no. 444.
- [17] M. A. Lefsky, A. T. Hudak, W. B. Cohen, and S. A. Acker, “Geographic variability in lidar predictions of forest stand structure in the Pacific northwest,” *Remote Sens. Environ.*, vol. 95, no. 4, pp. 532–548, Apr. 30, 2005.
- [18] V. Thomas, R. D. Oliver, K. Lim, and M. Woods, “LiDAR and Weibull modeling of diameter and basal area,” *Forestry Chronicle*, vol. 84, no. 6, pp. 866–875, 2008.

- [19] A. T. Hudak *et al.*, “Regression modeling and mapping of coniferous forest basal area and tree density from discrete-return lidar and multispectral satellite data,” *Can. J. Remote Sens.*, vol. 32, no. 2, pp. 126–138, Apr. 2006.
- [20] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [21] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [22] R. S. Olson, W. L. Cava, Z. Mustahsan, A. Varik, and J. H. Moore, “Data-driven advice for applying machine learning to bioinformatics problems,” *Pac. Symp. Biocomput.*, vol. 23, pp. 192–203, 2018.
- [23] G. J. Briem, J. A. Benediktsson, and J. R. Sveinsson, “Multiple classifiers applied to multisource remote sensing data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2291–2299, Oct. 2002.
- [24] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.
- [25] *The Living Murray Story: One of Australia’s Largest River Restoration Projects*. Canberra, ACT, Australia: Murray-Darling Basin Authority, 2011.
- [26] “Murray region: Barmah–Millewa forest,” Murray-Darling Basin Authority, Canberra, ACT, Australia, 2010.
- [27] J. A. Kerle, “Collation and review of stem density data and thinning prescriptions for the vegetation communities of New South Wales,” Dept. Environ. Conserv., Policy Sci. Division, 2005.
- [28] G. J. Horner, P. J. Baker, R. M. Nally, S. C. Cunningham, J. R. Thomson, and F. Hamilton, “Forest structure, habitat and carbon benefits from thinning floodplain forests: Managing early stand density makes a difference,” *Forest Ecol. Manage.*, vol. 259, no. 3, pp. 286–293, Jan. 25, 2010.
- [29] “Elevation information system,” ELVIS, 2020. [Online]. Available: <https://elevation.fsdf.org.au/>
- [30] J. Vaze *et al.*, “Floodplain inundation modelling for the Edward-Wakool region,” 2018.
- [31] “LASTools—Efficient tools for LiDAR processing,” rapidlasso GmbH, 2020. [Online]. Available: [http://lastools.org/download/lascanopy\\_README.txt](http://lastools.org/download/lascanopy_README.txt)
- [32] K. Y. van Ewijk, P. M. Treitz, and N. A. Scott, “Characterizing forest succession in central Ontario using lidar-derived indices,” *Photogramm. Eng. Remote Sens.*, vol. 77, no. 3, pp. 261–269, 2011.
- [33] R. O. Curtis and D. D. Marshall, “Technical note: Why quadratic mean diameter?,” *Western J. Appl. Forestry*, vol. 15, no. 3, pp. 137–139, 2000.
- [34] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [35] T. Mitchell and E. M. Munson, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
- [36] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.
- [37] L. S. Shapley, “A value for n-person games,” *Contrib. Theory Games*, vol. 2, no. 28, pp. 307–317, 1953.
- [38] I. Shendryk, M. Broich, and M. G. Tulbure, “Multi-sensor airborne and satellite data for upscaling tree number information in a structurally complex Eucalypt forest,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 73, pp. 397–406, Dec. 2018.
- [39] G. Y. Zhang and Y. Lu, “Bias-corrected random forests in regression,” *J. Appl. Statist.*, vol. 39, no. 1, pp. 151–160, 2012.



**Yuri Shendryk** (Member, IEEE) received the B.Sc. degree in geology from the Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, in 2009, the M.Sc. degree in geography from Lund University, Lund, Sweden, in 2013, and the Ph.D. degree in geography and remote sensing from the University of New South Wales, Sydney, NSW, Australia, in 2017.

He was a Postdoctoral Fellow with the Commonwealth Scientific and Industrial Research Organization, Canberra, ACT, Australia, focusing his research on the integration of airborne- and satellite-based remote sensing for forest monitoring, precision agriculture, and weed detection. In October 2020, he joined Dendra Systems (Australia) to apply his knowledge of GIS, remote sensing, and machine learning for large-scale ecosystem restoration.



**Emma Gorrod** received the B.Sc. degree in advanced environmental science from the University of New South Wales, Sydney, NSW, Australia, in 2002, and the Ph.D. degree in ecology from the University of Melbourne, Parkville VIC, Australia, in 2011.

She is currently a Principal Scientist of adaptive management with the Department of Planning, Industry, and Environment, Parramatta, NSW, Australia. She integrates empirical and modeled ecological data into management decisions for conservation outcomes. She is also an Adjunct Fellow with the Centre for Ecosystem Science, University of New South Wales, Sydney, NSW, Australia, and a Conjoint Fellow with the University of Newcastle, Callaghan, NSW, Australia.