

Integrating Commercial Thinning Estimation Research into the GeoHub Geospatial Intelligence Platform: An Application Example

Gulbe-Viluma Linda^{1,#}, Barga Ilze¹, Alexandru Andreea-Eliza¹, Grishanovs Andrejs¹

¹ Baltic Satellite Service, Ltd., Jurmala, LV-2015, Latvia, linda@forestradar.com, ilze@forestradar.com, eliza@forestradar.com, andrei@geohub.net

Abstract

The rapid evolution of artificial intelligence (AI) and geospatial information technologies is transforming how Earth Observation (EO) data are acquired, processed, and applied to address complex environmental and societal challenges. Despite the increasing availability of multi-source geospatial data, many organizations still face difficulties in converting these data into operational, decision-ready information. This study presents GeoHub, an integrated geospatial intelligence platform designed to transform EO data into actionable insights for environmental monitoring and forest management. The platform integrates satellite imagery from the Copernicus Programme, UAV observations, LiDAR datasets, in-situ sensors, and other geospatial layers within a unified analytical environment supported by AI-driven processing and analysis. One of the latest innovations is ML4HealthyForest service developed within the Interreg Estonia-Latvia program, which combines countrywide LiDAR data and machine learning to identify forest stands requiring thinning and estimate the need for both commercial and young stand thinning. The resulting service is available to forest owners via the GeoHub platform in Latvia and through the ForestMan web application in Estonia. This initiative demonstrates how an AI-enabled geospatial platform can convert complex EO data streams into accessible digital services that support sustainable forest management, climate resilience, and evidence-based environmental decision-making. Furthermore, Baltic Satellite Service demonstrates another Interreg Estonia-Latvia program's project Twinnovation, where high-frequency PlanetScope satellite imagery (3 m resolution) enables effective forest health monitoring and early detection of disturbances such as pest outbreaks, diseases, wildfire damage, and water stress.

Keywords: Earth Observation, machine learning, geospatial intelligence platform, forest monitoring, multi-sensor data integration

Introduction

The rapid growth of Earth Observation data and advances in artificial intelligence are reshaping geospatial information management and environmental monitoring. Large volumes of satellite imagery, LiDAR datasets, UAV observations, and in-situ sensor data are continuously collected, providing valuable information for understanding environmental change and managing natural resources. However, many organizations still face significant challenges in transforming these heterogeneous data sources into operational insights that can support timely decision-making. The complexity of data processing workflows, fragmented analytical tools, and limited interoperability between systems often slow down the practical use of geospatial intelligence.

This study investigates how an integrated geospatial platform can bridge the gap between raw EO data and practical decision-support services. The research analyses the design and application of GeoHub, an AI-enabled platform that integrates multi-source geospatial data, automated processing pipelines, and interoperable data services to support real-time environmental monitoring. The objective of this research is to evaluate machine learning model accuracy in directly predicting commercial forest thinning recommendations using LiDAR point cloud features and to present inclusion of the research based developments in GeoHub

platform for further usage. This study was a part of the Interreg Estonia-Latvia cooperation program initiative ML4HealthyForest. In addition, to the main objective, examples from another Interreg Estonia-Latvia cooperation program project Twinnovation are added to show GeoHub applicability to different data set up for detecting forest damages. ML4HealthyForest employs freely available data sets and provides results for the end-users through API while Twinnovation uses commercial satellite data and results are presented through web service.

Data and Method

Area of Interest

Both example projects ML4HealthyForest and Twinnovation focused on forested regions in the Baltic countries, particularly Latvia and Estonia, where large-scale digital forest monitoring services have been developed and implemented covering the whole area of the countries. The forests are predominantly mixed coniferous–broadleaf stands typical of the hemiboreal zone, with dominant tree species including Scots pine, Norway spruce, and deciduous species such as birch species, Gray alder, and Black alder. Despite the relatively small number of dominant tree species, forest stands exhibit high variability in species composition, management practices, and structural characteristics such as stand height and density.

Data Sets

Example projects represent two different data acquisition and processing approaches based on data accessibility. ML4HealthyForest used multi-source freely available geospatial datasets integrated within the GeoHub platform: countrywide LiDAR data combined with national forest inventory data, and information on cutting notices. This enables a high level of automation for end users, as all required data are already preprocessed and stored in the cloud. In contrast, forest damage detection within the Twinnovation project relies on a time series of commercial PlanetScope imagery, which must be ordered and downloaded on demand.

Methodology

Data processing and analysis were performed using the GeoHub geospatial intelligence platform, which integrates satellite data processing tools, machine learning models, and GIS-based analytical workflows. The whole process starting from raw data and ending with the required geospatial data set for each project can be divided in the following broad steps: 1) development of data processing/ML workflows, 2) validation of these workflows, 3) integration in the GeoHub, 4) providing the results for the end-users. The ML4HealthyForest example is presented below to illustrate these four steps.

Development of data processing/ML workflows

One of the workflows involved deriving features from LiDAR point clouds and training a machine learning model to directly predict the need for commercial and young stand thinning. Ground-truth preparation was strongly inspired by the work of Arumäe et al. (2022). Training and test polygons were selected using the Estonian cutting-notices database, which contains forest stand polygons for which cutting notices have been issued, along with the dates of notice issuance (rather than the dates when thinning operations actually began or ended). This dataset was complemented with stand age and dominant tree species information obtained from the Estonian Forest Registry. Attribute join was performed using spatial (GIS-based) joins, applying the mean value for stand age and the mode value for tree species within each polygon. Stand age was recalculated to the year 2024 based on the inventory year and recorded stand age from the Forest Registry. In addition, the available LiDAR data acquisition date was assigned to each polygon, enabling direct comparison with the cutting-notice dates.

Ground truth classes were assigned in the following way:

- 1: young stand; thinning not required - stand not older than 20 years, cutting notice year is from 0 - 2 years before lidar acquisition.
- 2: young stand; thinning required - stand not older than 20 years, cutting notice year is from 0 - 2 years after lidar acquisition.
- 3: stand at commercial thinning age; thinning not required - stand older than 20 years, cutting notice year is from 0 - 2 years before lidar acquisition.
- 4: stand at commercial thinning age; thinning required - stand older than 20 years, cutting notice year is from 0 - 2 years after lidar acquisition.

LiDAR feature selection was based on the assumption that the need for thinning can be estimated by analyzing LiDAR point density across different height layers above ground. Unlike the methodology used in Arumäe et al. (2022), we calculated features using a grid-based approach that is independent of stand boundary maps. This approach provides the advantage of allowing features to be precomputed and stored for the entire Estonia and Latvia, enabling rapid estimation for any stand submitted by an end user, requiring only the stand boundaries of interest.

For each stand in the ground truth vector, the mean raster values of the feature images generated in the previous step were calculated to create the feature array. Target values were extracted directly from the attribute table of the ground truth data. A Random Forest classifier from the scikit-learn Python library was employed as the machine learning model, using 21 estimators and 4 independent models were fitted for 4 dominant tree species: Scots pine, Norway spruce, birch and black alder.

Validation of the workflows

Validation workflow design typically is determined by the available data sets or resources for field inspections. ML4HealthyForest relied on a traditional validation workflow using an independent test set separated from the initial ground truth data while the Twinnovation project included additional field inspection based validation done by end-users in their particular forest stands.

Validation procedures for workflows integrated in GeoHub uses a well known confusion matrix with its derived metrics like overall accuracy (OA) with an emphasis on explaining True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) cases.

For example, thinning recommendation accuracy was evaluated for 4 dominant tree species separately by using 1400 test stands randomly selected from the whole area of Estonia. If a model is evaluated on only one train/test split, results can depend heavily on how the data was split. Cross-validation reduces this variability. The dataset is divided into several folds, and the model is trained on all but one fold and tested on the remaining fold; this process is repeated so that each fold is used once for testing. The function returns a score for each split, giving a more reliable estimate of how well the model generalizes to unseen data by reducing the dependence on any single, potentially lucky or unlucky, split. We evaluated 15 folds for each tree species.

Integration in the GeoHub and providing data sets to end-users

Once the data processing and machine learning workflows were developed as Python tools, they were integrated into the GeoHub platform. Data storage and management are coordinated by the platform, as several applications and research projects rely on the same datasets. To simplify usage for end users, data features are pre-calculated and stored for the entire Baltic States region whenever possible. For example, features derived from LiDAR data are computed

in advance. When an end user uploads their forest stands of interest, the relevant features are retrieved from storage, passed to the machine learning workflow, and the results are returned to the user through an API. Data can also be visualized through the GeoHub web service, ensuring accessibility without requiring additional software installation on the end-user side. Fig 1 shows data samples for thinning estimation response and forest damage detection results.

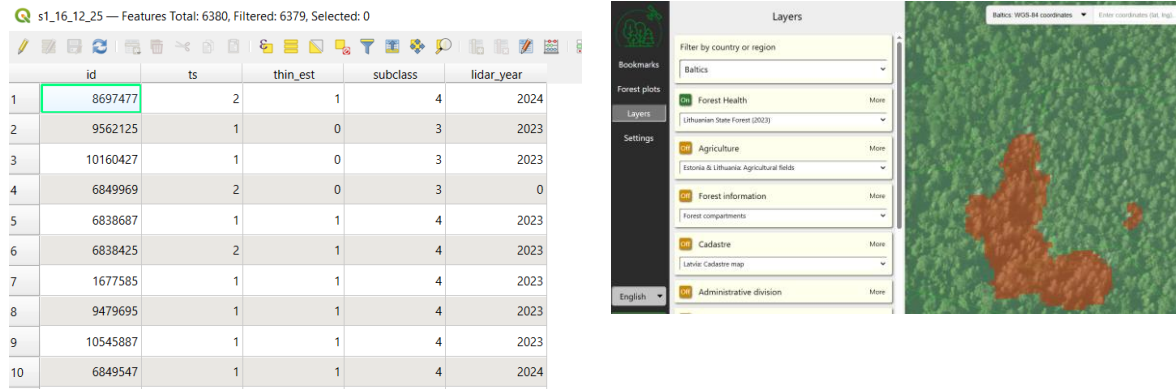


Fig. 1. Data for end users are available through 1) API service like image to the left shows example of ML4Healthy project output where user uploaded stand polygons are complemented with thinning decision field thin_est, 2) web service like shown in the image to the right, where damaged forest area is delineated and visualized over RGB ortophoto.

Results

Within the ML4HealthyForest Project, the thinning estimation model presented here achieved 61–76% accuracy depending on dominant tree species, see Table 1.

Table 1. Accuracy assessment result for 1400 tests stands analysed in ML4Healthy Project.

Dominant tree species in stand	Average OA in 15 folds	Standard deviation of OA in 15 folds
Scots pine	0.74	0.01
Norway spruce	0.76	0.02
Birch species	0.61	0.01
Black alder	0.70	0.02

Analysis of TP, TN, FP and FN cases showed strong dependency on LiDAR point density in the first 9 meters above the ground. TP cases tend to have smaller point densities above the ground compared to TN cases. In addition, the number of FN cases was relatively small and typically caused by roads going through the test polygons or other gaps in the canopy. However, the number of FP cases was high. Statistical analysis in boxplots indicated that FP cases occur in scenarios where lidar point density is lower in the first 3 meters above ground indicating that a lot of pulses cannot penetrate the canopy. More information of the actual situation in the field would be required to understand those FP cases, for example, those can be actual TP due to delayed work or maybe different thinning practices need to be incorporated in the model.

Validation results indicate that further improvements can be achieved by integrating LiDAR- and inventory-based workflows, transitioning from stand-based to grid-based analysis also in the result acquisition phase, and using updated LiDAR datasets.

Furthermore, the Twinnovation Project demonstrates the value of high-frequency PlanetScope satellite imagery (3 m resolution) for forest health monitoring, enabling early detection of disturbances such as pest outbreaks, disease, wildfire damage, and water stress with achieved

92-95% accuracy. Accuracy was assessed during field inspections: totally 62 forest stands were checked where 52 confirmed damage, while 10 FP cases were identified mostly due to cutting activities. After introducing context evaluation the number of false positives reduced from 16% to 5%.

The results demonstrate that integrating multi-source EO data within the GeoHub platform enables efficient large-scale forest monitoring and decision-support services. The combined use of satellite imagery, nationwide LiDAR datasets, and machine learning models allows the identification of forest stands requiring thinning and supports the assessment of forest health risks across large forest areas in the Baltic States.

Conclusion

This study presented research results on estimating the need for commercial and young stand thinning directly from LiDAR point cloud features using Random Forest classifiers for four dominant tree species. In addition to the research findings, the study described integration principles within the GeoHub platform and provided additional examples from the Twinnovation project that demonstrate different aspects of using the platform.

The accuracy of the data processing workflows ranged from 61% to 76% and continues to be improved through ongoing research. However, the integration within the GeoHub platform enables easy updates to the workflows and facilitates testing with end users already during the development phase.

The ML4HealthyForest and Twinnovation projects confirm that AI-enabled EO services can be successfully operationalized for practical use by private forest owners and environmental stakeholders, as demonstrated through the GeoHub and ForestMan platforms.

References

Planet Labs (2025). PlanetScope. <https://docs.planet.com/data/imagery/planetscope/>

Arumäe, T., Lang, M., Sims, A., & Laarmann, D. (2022). Planning of commercial thinnings using machine learning and airborne lidar data. *Forests*, 13(2), 206.