Miti360: An integrated dataset combining remote sensing, ground measurements and weather data for improved reforestation monitoring

Technical Report on Dataset Collection and Curation

Cedric Kiplimo[1], Samuel Mbatia[1], Viola Orina[2], Ciira wa Maina[1], Arthur Sichangi[2], Denis Gitundu[2]

[1]Centre for Data Science and Artificial Intelligence (DSAIL), Dedan Kimathi University of Technology, Nyeri, Kenya

[2]Institute of Geomatics, GIS and Remote Sensing (IGGReS), Dedan Kimathi University of Technology, Nyeri, Kenya

Table of Contents

1 Background

In today's age of artificial intelligence (AI), machine learning (ML) has the potential to improve forest monitoring efforts when coupled with appropriate data such as remote sensing data and ground measurements [1], [2], [3]. The application of these tools to analyse satellite and drone imagery can speed up the accurate estimation of biomass and individual tree monitoring. Despite the huge interest in applying ML to forestry, the scarcity of machine learning ready datasets related to forestry in Africa remains a major hindrance [3]. The major datasets include NEON Crowns [1] and the Auto Arborist dataset [2] with data from North America, ReforesTree [4] with data from Ecuador, and a new dataset from Northern Australia [5]. This project was undertaken to address the existing data gap by collecting, annotating, and providing access to information that supports data-driven decision-making in establishing reforested stands and monitoring the success of reforestation efforts. Miti360 will add to the current collection of ML-ready datasets in forestry, which currently have limited geographic diversity.

In practice, data on forest resources is usually collected systematically through a process called forest inventory. This process is crucial for understanding forest composition, structure, and health, thereby supporting effective management and conservation efforts. For this project, the focus of the forest inventory was on gathering detailed measurements of individual tree biophysical parameters, precise tree locations, and species identification. Furthermore, weather data from stations located within a 100 km radius of the study site were collected to enable analysis of how early growth phases—an especially critical period for reforested stand establishment—respond to varying weather conditions. Such a comprehensive dataset that brings together accurate ground data, terrestrial (stereoscopic and single) and remote sensing imagery, and weather data will facilitate forest monitoring research efforts in ways that current publicly available datasets cannot.

Accurate forest monitoring, especially at the individual tree level, will be immensely beneficial to African countries where forest inventories still lag those of developed nations in terms of scale, accuracy and reliability [6]. In Kenya, for example, current efforts to plant up to 15 billion trees by 2032 present the unique challenge of monitoring the growth of those individual trees to ascertain their survival and assess their growth rates. These are challenges that Miti360 is uniquely positioned to help solve by facilitating the development of the relevant machine learning models. Similar efforts have been undertaken in the Global North [2], [7], [8], [9].

To harness the full potential of data-driven forestry and address pressing challenges in reforestation monitoring, advanced machine learning techniques are being integrated into research and practical field applications. In particular, the machine learning algorithms trained using Miti360 will be able to, among other things:

i. Identify and classify species of individual young trees, saplings and seedlings in heterogeneous stands from drone images,
ii. Estimate biophysical parameters such as crown diameter and tree height and from these derive biomass estimates, and
iii. Predict changes in tree biophysical parameters and stand volume in relation to prevailing weather conditions and tree species.

We believe that this will make accurate quantitative analysis of reforested stands possible and this will in turn promote reforestation efforts.

2 Methods

2.1 Study Area

The study area selected for this project was a reforested stand within the Kieni Forest in Kenya (Figure 1). This forest is situated within the Aberdare Ecosystem, one of Kenya's five main water towers which supplies 80% of the water used in Nairobi. The entire ecosystem spreads across four counties—Kiambu, Murang'a, Nyeri, and Nyandarua—with Kieni Forest located within Kiambu. Kieni Forest lies between 2200 m and 2684 m above sea level and receives rainfall of 1150 mm to 2560 mm annually [10]. The long rain season stretches from March to June, while short rains are received from October and December. Its soils are rich in organic matter, making them fertile and favourable to developing thick undergrowth. Its vegetation comprises natural forests, plantations, bamboo, meadows, and tea zones. The replanted section covers an area of 770 ha and contains more than ten indigenous species, mainly *Dombeya torrida*, *Juniperus procera*, *Olea africana*, and *Prunus africana*. At the beginning of this data collection, the crown diameters range from 0.35 m to 6.6 m, the heights from 0.9 m to 7.2 m, and the basal diameters from 2.2 cm and 32.8 cm. Being a reservoir of biodiversity, Kieni Forest also has a wide range of fauna, such as the African elephant (*Lexodonta africana*), duiker (*Neottragus moschatus*), Bush pig (*Patomochoerus porcuso*), and mongoose (*Helogale parvula*), among others [10], [11].

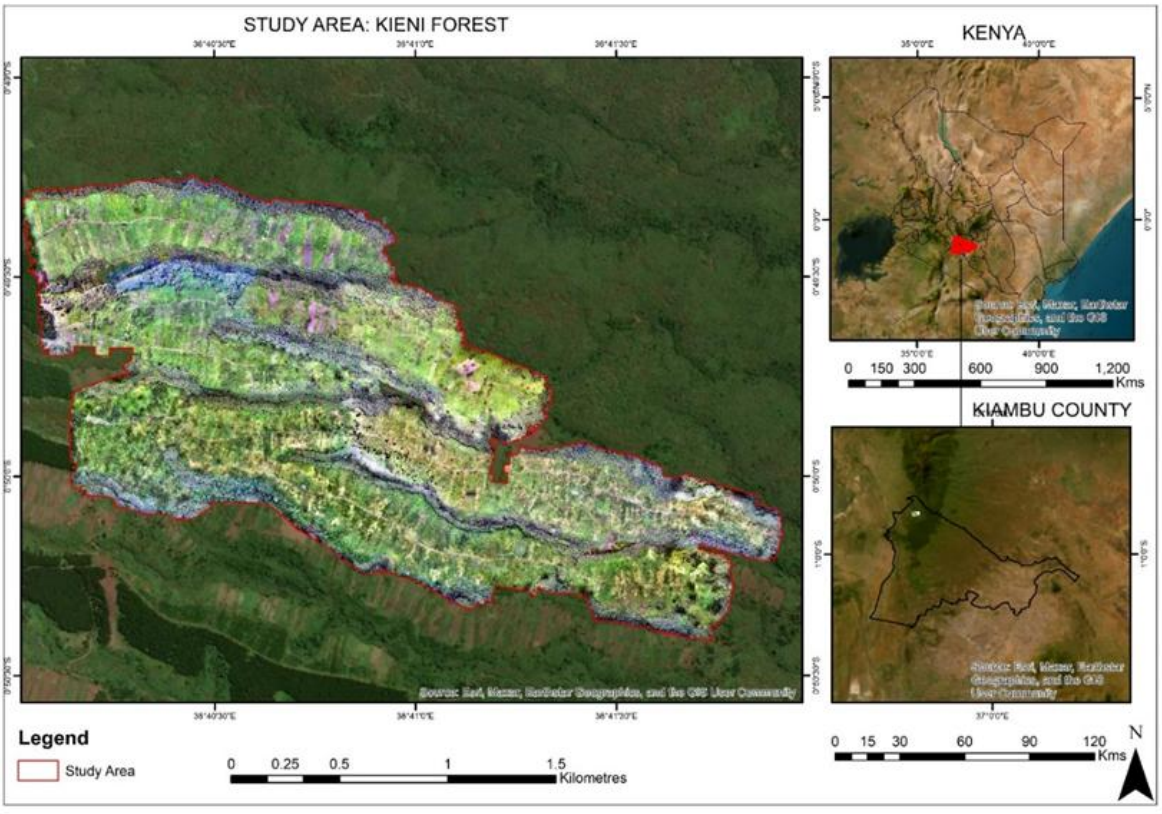

Figure 1: The study area - a reforested section of Kieni Forest

2.2 Materials

The stereo images were captured using an inhouse stereo camera comprising two Logitech C270 USB web cameras, which have a resolution of 720 × 1280 pixels, in an assembly with a baseline of 12.9 cm. This stereo camera was powered by the TreeVision software [12] running on an NVIDIA Jetson Nano 2 GB Developer Kit and operated via a connected HDMI mini screen. The features of this kit include a quad-core ARM A57 @ 1.43 GHz processor, 4 GB 64-bit LPDDR4 RAM, and a 128-core Maxwell GPU. The single images were captured using an Oppo F11 smartphone with a 3840 × 2160 camera sensor. To capture the GPS coordinates of each tree, the Garmin GPSMAP 64s handheld outdoor GPS was used. Finally, the ground truth biophysical parameters were taken using a tape measure and a graduated height pole.

2.3 Field Data Collection

2.3.1 Sampling Design

For this project, a systematic sampling approach was selected for carrying out the forest inventory exercise (Figure 2). In systematic sampling, sample plots are spaced at fixed intervals throughout the entire population. This sampling method was chosen because it provides a good balance between statistical rigour and operational feasibility in the field. Additionally, it is compatible with remote sensing data such as gridded raster data, thus making geostatistical modelling with collected data possible. The study area shown in Figure 2 was divided into contiguous units measuring 100 m × 100 m. Sampling plots were then selected systematically from these units such that the distance between consecutive plots was 100 m horizontally and 200 m vertically. In Figure 2, the shaded units are the sampling plots. The sampling design yielded a total of 56 plots. Of these, trees were sampled from only 41 while the rest were labelled as invalid plots because they were either covered entirely by bamboo bushes or they had no trees. From each plot, the plan was to sample 15 trees randomly to make a total of 615 trees.

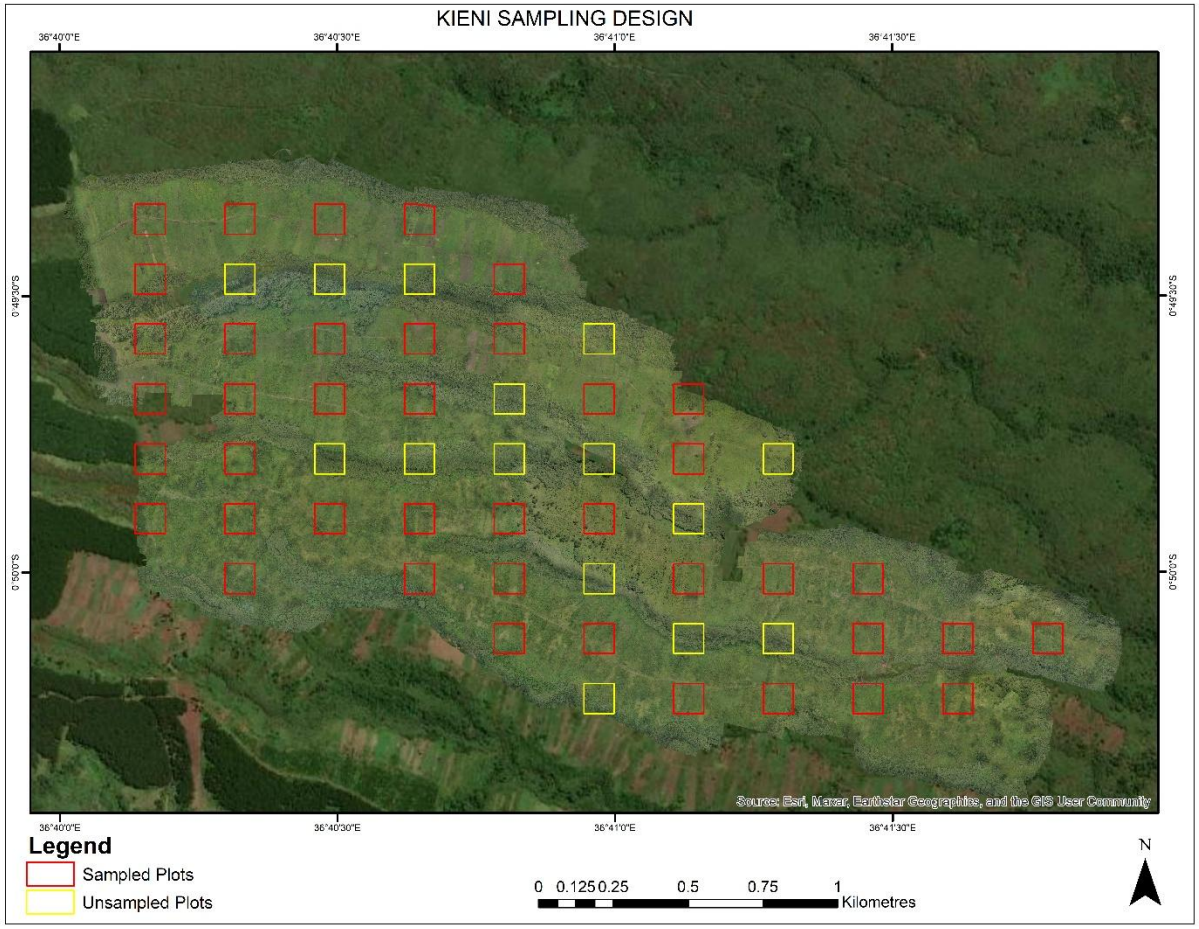

Figure 2: Sampling design

2.3.2 Ground Truth Data Measurement & GPS Data Collection

The tree biophysical parameters measured during the tree sampling were the tree height (TH), crown diameter (CD), and basal diameter (BD). These were measured as follows:

- Tree height: The height of each tree was measured using a graduated pole. The pole next to the tree such that it stood parallel to the tree with its base was at the same level as the tree trunk's base and its height was read off from the pole.
- Crown diameter: To measure the diameter of each crown, a tape measure was stretched along an axis stretching from one edge, through the crown's centre to the opposite edge and the reading recorded. Although the standard practice in forestry is to take the average diameter along two perpendicular axes, we chose to use a single axis due to practical considerations. However, these measurements can be augmented with those extracted from georeferenced orthophotos which have been shown to be more accurate.
- Basal diameter: This measurement was taken by wrapping a tape measure around the tree trunk at between 5 cm and 10 cm above ground and then computing the diameter from the circumference.

The location of each tree was recorded using the Garmin GPSMAP 64s handheld outdoor GPS, which has an accuracy of 3 m.

2.3.3 Terrestrial Image Collection

For each tree sampled in the study area, two kinds of terrestrial images were captured – stereoscopic (paired left and right) images and single images. The stereo images were captured using the inhouse stereo camera described in section 2.2 while the single images were taken using the smartphone described in the same section. In both cases, the images were taken such that the tree is the main subject of the image and is positioned at the centre of the frame.

2.4 Drone Survey

Planning for the drone survey involved two main steps. First, the reforested area was divided into survey blocks that were small enough to be fully covered using a single drone battery. Second, flight plans were prepared for each block using consistent flight parameters.

2.5 Survey Block Design

The Area of Interest (AOI) was outlined by digitizing the reforested sections of Kieni Forest using a high-resolution base satellite image in ArcGIS. A grid of square quadrats measuring 150 meters was then overlaid and intersected with the digitized AOI. The quadrats were deliberately kept small to allow for merging rather than splitting during the survey design. Adjacent intersected quadrats were merged to create the final survey blocks. Each block was limited to a maximum area of 0.2 square kilometers, which was suitable for a single drone battery when flying at approximately 90 meters altitude with a flight path spacing of about 30 meters. These parameters were informed by previous flight tests. In total, 21 survey blocks were generated, each measuring between 0.08 and 0.18 square kilometers in size (Figure 3Figure 2).

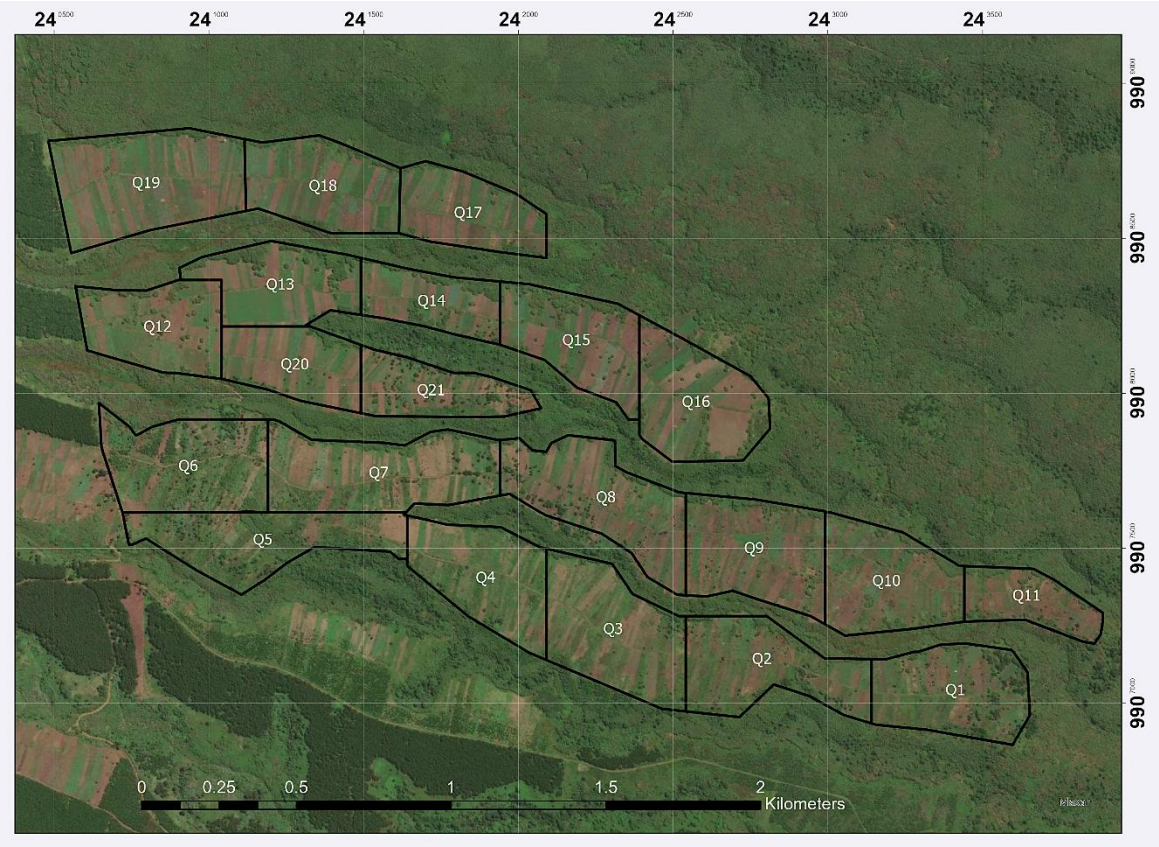

Figure 3: Final Survey Blocks, numbered according to the order of data collection.

2.6 Flight Planning

Flight plans were developed for each survey block using DJI Flight Planner software. A ground sampling distance (GSD) of 2 cm was targeted, resulting in a flying altitude of approximately 88 meters (288 feet) above the take-off point. To ensure sufficient image overlap for high-quality orthomosaic generation, a forward overlap of 80% and a side overlap of 70% were applied. These parameters yielded a flight path spacing of approximately 32 meters. The final flight plan (Figure 4) depicts the full mapped area, survey block, flight paths, and the individual camera exposure points.

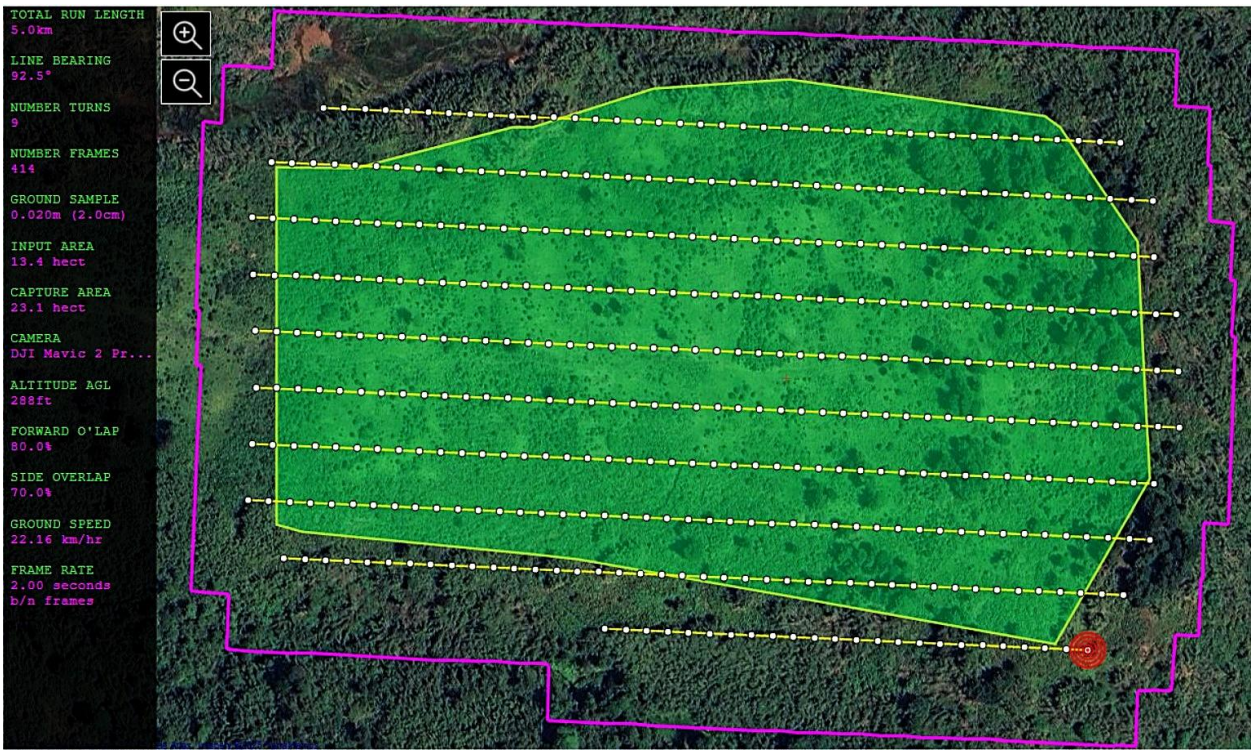

Figure 4: Sample flight plan showing the full mapped area (purple outline), survey block (green polygon), flight paths (yellow lines), and individual camera exposure points (white dots)

2.7 Drone Image Processing

2.7.1 Image Alignment

The processing and georeferencing of the drone images were done using Agisoft Metashape, a leading photogrammetric software widely used for Structure-from-Motion (SfM) processing. A total of 4631 images were captured during the first phase of the data collection and 4600 images during the second phase. These images had a 70-80% forward overlap and a 60% side overlap. The overlap is critical for photogrammetric processing, as it allows Agisoft Metashape to detect matching key points across multiple images. The drone's onboard GPS recorded approximate image positions, which provided initial georeferencing data for image alignment and subsequent processing.

Once the captured images were imported into Agisoft Metashape, the software automatically reads the embedded metadata (GPS coordinates, focal length, and camera parameters) from each image. To align the images, the software identified key feature points across overlapping images and reconstructed a sparse point cloud. This step estimated the camera positions and orientations, forming the initial 3D map of the study area.

2.7.2 Accuracy Enhancement with Ground Control Points

To improve geospatial accuracy, Ground Control Points (GCPs) were incorporated into the model. GCPs are precisely surveyed locations with known coordinates, typically collected using a high-precision GNSS receiver (e.g., RTK-GPS). These points were manually marked in multiple drone images within Metashape, and their real-world coordinates were inputted. The software then optimised the camera alignment and 3D model to match the GCPs, significantly reducing errors from drone GPS drift or lens distortion.

2.7.3 Dense Point Cloud Generation and Digital Elevation Model (DEM) creation

Following GCP optimisation, a dense point cloud was then generated. Agisoft Metashape does this using advanced multi-view stereo (MVS) algorithms to produce a highly detailed 3D representation of the terrain and vegetation [13]. From the dense cloud, a Digital Elevation Model (DEM) was derived, which served as the foundation for orthophoto rectification. The DEM ensured that elevation variations were accounted for, eliminating distortions in the final orthophoto.

2.7.4 Orthophoto Generation and Export

Once the DEM was ready, the final step was to generate the orthophoto. This process involved projecting the stitched images onto the DEM, correcting for perspective and topographic distortions. The resulting orthophoto was a seamless, georeferenced, and high-resolution image suitable for GIS applications. The output was exported in GeoTIFF format, preserving geospatial metadata for further analysis in software such as QGIS or ArcGIS.

2.8 Image Annotation

2.8.1 Terrestrial Images

For the terrestrial images (both stereo and single) captured in the study area, two labels were created – a semantic segmentation mask and a species ID. The masks were created using the Labelbox platform, which provides a suite of annotation tools including AI-assisted labelling. Each mask was created through a two-step process. First, AI-assisted labelling powered by the segment anything model (SAM) was used to form a preliminary mask. This was followed, if necessary, by edits to the preliminary mask to make it more accurate. The species of each tree was identified after visual inspection of the image. The annotators could do this with good accuracy after adequate training on species identification had been carried out.

2.8.2 Orthophotos

As part of the workflow, we conducted detailed data annotation and labelling on orthophoto tiles derived from drone-captured imagery of Kieni Forest. The orthophoto was divided into manageable tiles to facilitate precise annotation of individual tree crowns using Label Studio [14], an open-source data labelling tool. Each tile was carefully examined, and bounding boxes were manually drawn around tree crowns to demarcate their extents. The annotations were saved in JSON format, which stores both the geometric coordinates of the bounding boxes and associated metadata, such as class labels (e.g., tree). The use of Label Studio streamlined the annotation process by providing an intuitive interface for drawing, editing, and exporting labelled data. Quality control measures, including cross-validation between annotators and iterative reviews, were implemented to minimize errors.

2.9 Quality Control

2.9.1 Post-collection Verification

At the end of each day during the data collection exercises, all the data collected from the field were consolidated and checked to ensure there were no missing attributes (e.g., GPS coordinates or measurements) and that they were free from errors (e.g., corrupted images). This step prevented potential errors that often arise when data is accumulated, or significant time passes before verification.

2.9.2 *Image Annotation Training*

After the first data collection exercise was completed, a team of annotators underwent comprehensive training on how to perform high-quality image annotations using various tools commonly used in the industry. The training curriculum included an overview of computer vision and the need for image labels, overview of image annotation, tool-specific tutorials, and annotation tasks to be completed by the trainees and assessed by the trainers. This approach to training ensured that the annotators developed adequate skills and could produce high-quality labels.

2.9.3 *Task Review and Consensus Scoring*

The annotators were placed in teams, and each team was assigned a reviewer to ascertain the quality of submitted annotations. The workflow was set up in Labelbox such that every annotation submitted was reviewed before being marked as complete. Low quality annotations were rejected and sent back to the annotators for rework.

Labelbox provides a consensus scoring feature that is useful for analysing the quality of annotations submitted by annotators. This feature was set up by requiring at least two annotators to submit a mask for an image so that a consensus score would be computed based on the overlap between them. Annotations with consensus score of less than 80% were either rejected by reviewers and reworked or, where one of the masks was visibly of high quality, the good mask was accepted.

In the case of bounding box annotations for the orthophotos, quality control measures were similar and included cross-validation between annotators and iterative reviews to minimize errors.

2.9.4 *Final Inspection of Dataset*

Once the dataset curation was complete, all its components underwent a thorough inspection to ascertain that all necessary components were present and were of high quality. These final quality checks were done by randomly sampling from the dataset.

3 Results

3.1 Summary of Collected and Labelled Data

The data collected in this study consists of drone and ground-based images, individual tree attributes and aggregated data from TAHMO [15] weather stations. These are captured in Table 1 below.

Table 1: Summary of the dataset

#	Data Category	Data Type		Quantity	Format
1	Drone Images	Orthophoto		2	TIF
		Tiles		844	TIF
		Tree crown annotations		24000	JSON

		Tree crown species		1208	JSON
		Tree species shapefile		1208	SHP
2	Tree ground measurements	Numeric data		1208 (600 trees 2 times in 2024 - 2025)	JSON
3	Ground based single images	Images and tree masks		1208 (600 trees 2 times in 2024 - 2025)	JPEG
4	Tree stereo images	Images and tree masks		2416 (600 trees 2 times in 2024 - 2025)	JPEG
5	Weather data from 40 stations	Time series data		8 years daily data	API endpoint

Figure 5 and Figure 6 show sample image types from the dataset together with their annotations. These figures show the image masks for both terrestrial images as well as the bounding boxes around individual trees on the orthophoto tiles.

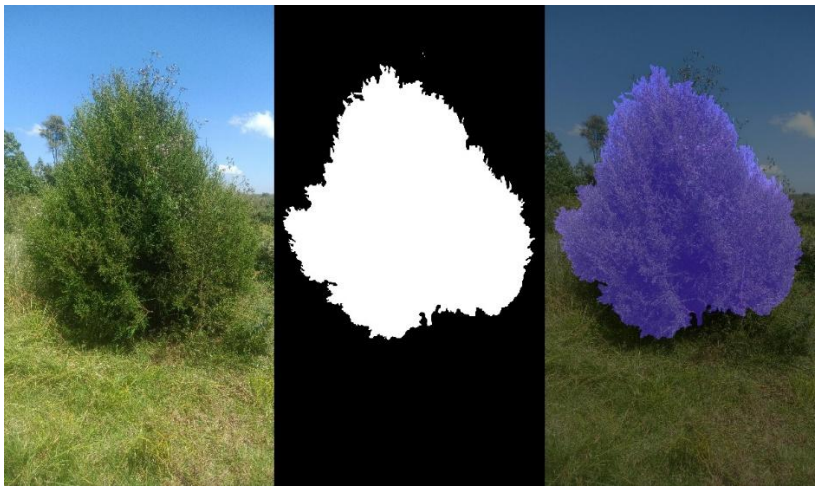

Figure 5: Sample terrestrial image from the dataset together with its annotation

Figure 6: Sample orthophoto tile from the dataset together with its annotations

3.2 Spatial Distribution of Sampled Trees

The spatial distribution of the sampled trees in the study area is shown in Figure 7.

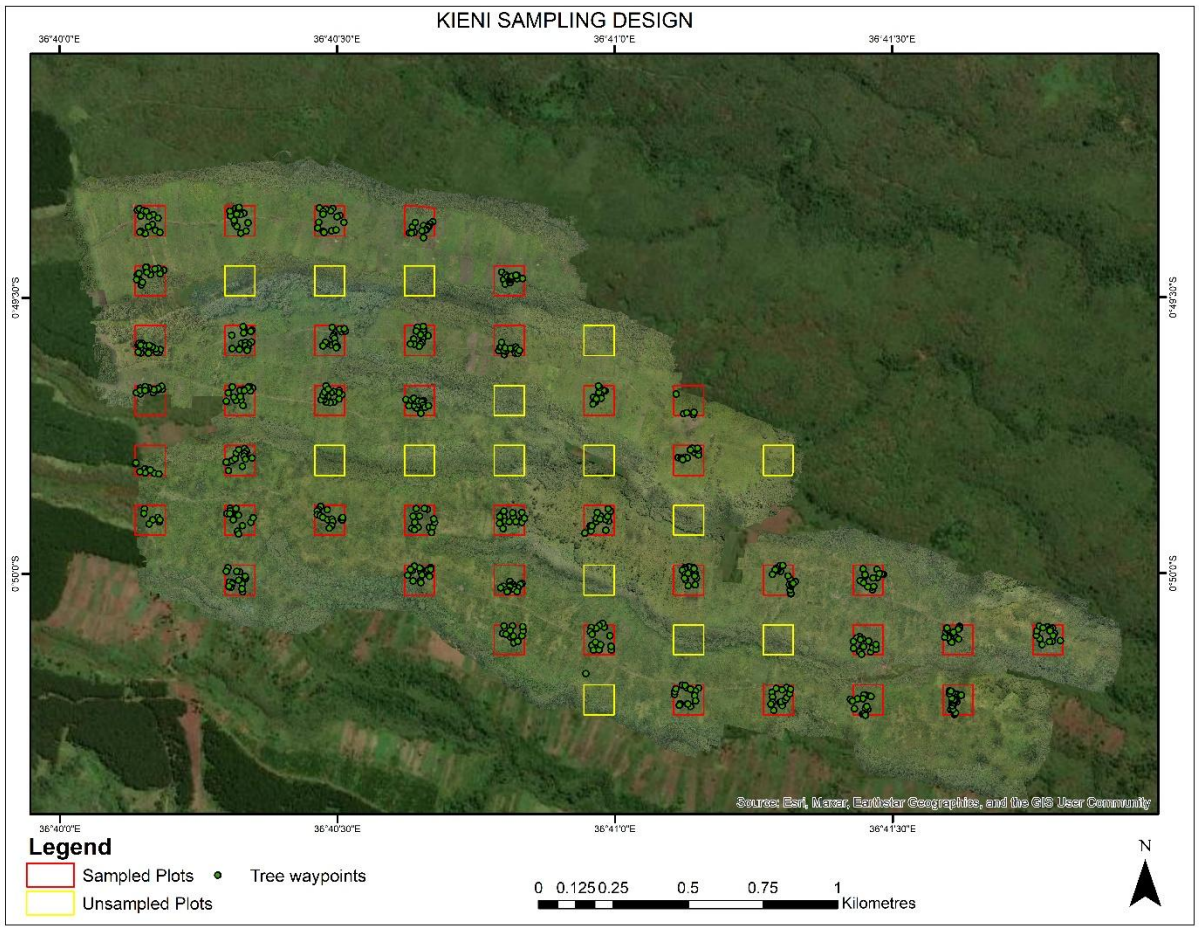

Figure 7: Spatial distribution of sampled trees

3.3 Distribution of Measurements

The distribution of the tree heights, crown diameters, and basal diameters as well as their joint distributions are shown in Figure 8 and Figure 9 respectively.

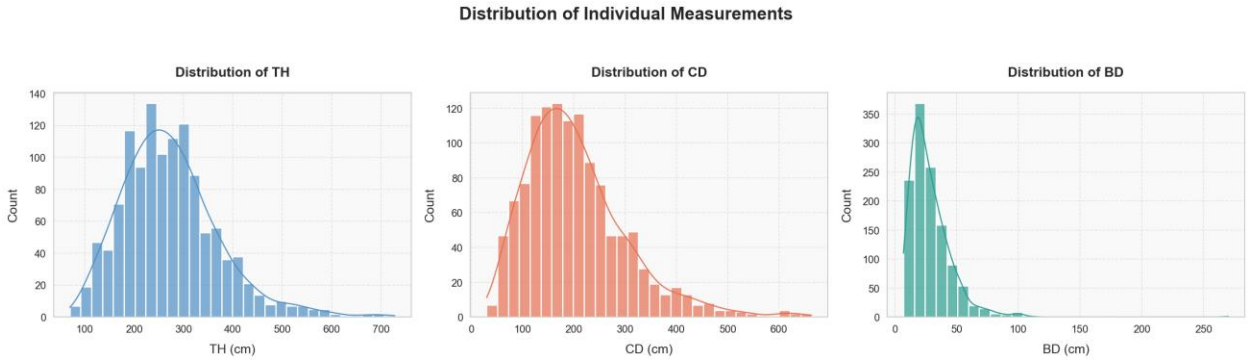

Figure 8: Distribution of individual tree biophysical parameters

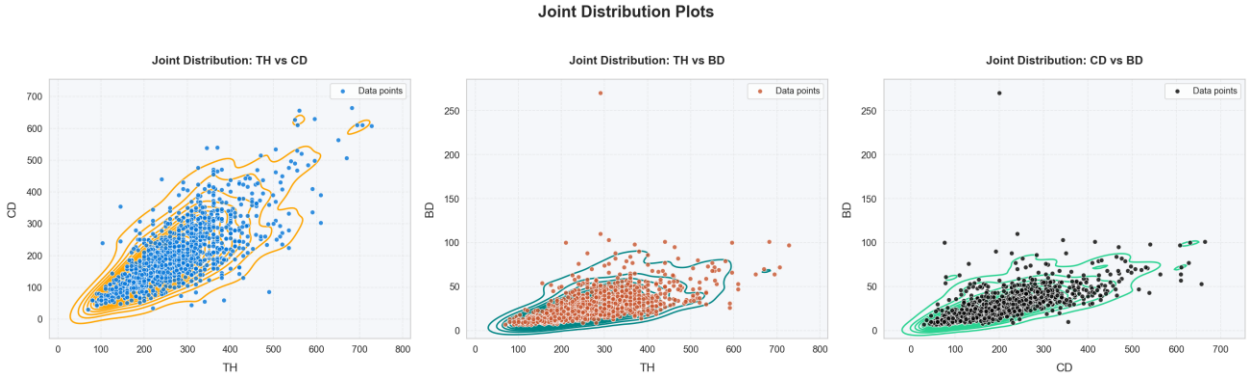

Figure 9: Joint distributions of the various tree biophysical parameters

4 Applications and Usage Guidelines

As various stakeholders engage in reforestation efforts, a key question is which trees should be planted, where they should be planted and when they should be planted. In preliminary work aimed at monitoring indigenous tree species at a 770-ha reforested stand in Kieni, Kenya [16], as well during our data collection exercises, we observed that some strategies employed include planting a wide variety of species with the hope that some would succeed. Miti360 will be useful in developing systems that power data driven decisions for stakeholders involved in the establishment and the maintenance of reforested stands. In many of Kenya's forests, plantations are often established with great assistance from farmers who are allowed to grow food crops within recently established stands. Our data will validate the contribution of these local farmers to reforestation efforts.

4.1 Appropriate Usage

Miti360 can be used in varied ways to train and assess machine learning models. One example is to build models that can detect individual trees in the presence of dense shrubs that make it difficult for human

beings to tell trees apart from shrubs. The labels of trees in the orthophotos will be useful for achieving that goal. One useful research angle we have pursued in the past is that of automating tree inventory using stereoscopic photogrammetry [12], [16]. With recent advances in deep learning and 3D computer vision, the stereoscopic images in Miti360 would be invaluable in developing better techniques for achieving the same goals. Regardless of the ways in which dataset may be used, we believe that all efforts directed towards developing novel techniques for forest monitoring tailored towards our African context will produce the greatest impact.

Overall, this data will serve stakeholders by enabling the development of machine learning systems capable of:

1. Increasing the efficiency of tree monitoring operations by quickly processing data collected by drones to determine individual tree counts.
2. Determining which species are growing well in a given area. This will be demonstrated by determining change in tree biophysical parameters over a period of one year.
3. Determining quantitative relationships between tree growth and weather.

4.2 Data and Code Availability

The labelled dataset and all associated catalogues are available for download from Zenodo. The source code used to analyse the data will be made available in due course.

4.3 Acknowledgements

This work was carried out with support from Lacuna Fund, and Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ).

The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute.

5 References

[1] Ben. G. Weinstein *et al.*, "NEON Crowns: a remote sensing derived dataset of 100 million individual tree crowns," Sep. 09, 2020. doi: 10.1101/2020.09.08.287839.

[2] S. Beery *et al.*, "The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 21262–21275. doi: 10.1109/CVPR52688.2022.02061.

[3] A. Ouaknine, T. Kattenborn, E. Laliberté, and D. Rolnick, "OpenForest: a data catalog for machine learning in forest monitoring," *Environmental Data Science*, vol. 4, p. e15, Feb. 2025, doi: 10.1017/eds.2024.53.

[4] G. Reiersen *et al.*, "ReforesTree: A Dataset for Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12119–12125, Jun. 2022, doi: 10.1609/aaai.v36i11.21471.

[5] A. J. Jansen *et al.*, “Deep Learning with Northern Australian Savanna Tree Species: A Novel Dataset,” *Data (Basel)*, vol. 8, no. 2, p. 44, Feb. 2023, doi: 10.3390/data8020044.

[6] K. Nesha *et al.*, “An assessment of data sources, data quality and changes in national forest monitoring capacities in the Global Forest Resources Assessment 2005–2020,” *Environmental Research Letters*, vol. 16, no. 5, p. 054029, May 2021, doi: 10.1088/1748-9326/abd81b.

[7] V. Grondin, J.-M. Fortin, F. Pomerleau, and P. Giguère, “Tree detection and diameter estimation based on deep learning,” *Forestry: An International Journal of Forest Research*, vol. 96, no. 2, pp. 264–276, Mar. 2023, doi: 10.1093/forestry/cpac043.

[8] Q. Ou, X. Lei, and C. Shen, “Individual Tree Diameter Growth Models of Larch–Spruce–Fir Mixed Forests Based on Machine Learning Algorithms,” *Forests*, vol. 10, no. 2, p. 187, Feb. 2019, doi: 10.3390/f10020187.

[9] G. D. Pearse, A. Y. S. Tan, M. S. Watt, M. O. Franz, and J. P. Dash, “Detecting and mapping tree seedlings in UAV imagery using convolutional neural networks and field-verified data,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 156–169, Oct. 2020, doi: 10.1016/j.isprsjprs.2020.08.005.

[10] J. Nyukuri, “Issues Influencing Sustainability of the Aberdare Range Forests: A Case of Kieni Forest in Gakoe Location, Kiambu County,” University of Nairobi, Nairobi, Kenya, 2012.

[11] Kenya Forest Service, “Aberdare Forest Reserve Management Plan,” Nairobi, Kenya, 2010.

[12] C. Kiplimo, C. E. Epege, C. wa Maina, and B. Okal, “DSAIL-TreeVision: A software tool for extracting tree biophysical parameters from stereoscopic images,” *SoftwareX*, vol. 26, p. 101661, May 2024, doi: 10.1016/j.softx.2024.101661.

[13] Agisoft, “Metashape 2.2: Photogrammetry + LiDAR,” 2025, *St. Petersburg*.

[14] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, “Label Studio: Data labeling software,” 2025.

[15] N. van de Giesen, R. Hut, and J. Selker, “The Trans-African Hydro-Meteorological Observatory (TAHMO),” *WIREs Water*, vol. 1, no. 4, pp. 341–348, Jul. 2014, doi: 10.1002/wat2.1034.

[16] C. Kiplimo, C. wa Maina, and B. Okal, “Low-Cost Non-Contact Forest Inventory: A Case Study of Kieni Forest in Kenya,” *Challenges*, vol. 15, no. 1, p. 16, Mar. 2024, doi: 10.3390/challe15010016.