Multi-Sensor Detection Methodology for Conservation Monitoring Based on Artificial Intelligence

Mata Elang Platform - Case Study of Baluran National Park, Indonesia
Interim Methodology Report

From the air, analysed by machine, decided by people

Badar Mubarok Yogaswara
Version 1.2 (interim) - September 2026

mataelang.elps.co.id

Document Status Statement

CAUTION: This is an INTERIM REPORT. The system described is at proof-of-concept stage with real pilot data; it is NOT a fully validated system. Performance figures come from small validation sets that do not support claims of generalisation. This document must not be cited as evidence of operational readiness, and system outputs must not serve as the sole basis for legal, enforcement, or policy decisions.

Every known limitation is stated openly in Chapter 7. The author holds that an honest interim report serves science and policy better than a final report that conceals its uncertainty.

Note: To readers not yet familiar with the terminology of monitoring based on artificial intelligence: please read Appendices A to C - Chapters 13 to 15 - first. The table in Section 13.1 sets out reader profiles and which appendix suits each. Those three appendices explain the foundations from the beginning and alter nothing in Chapters 1 to 12 - readers already familiar with the field may skip them entirely.

Beginning with the appendices is not a shortcut but the fastest route to being able to ASSESS this report rather than merely believe it. This document was written to be assessed. Its readers are entitled to know not only what is claimed but how strong the basis for each claim is - and Appendix C in particular prepares the reader to test this work exactly as they would test anyone else's.

1. Introduction

1.1 Background

Conservation monitoring in Indonesia faces a fundamental mismatch between the area that must be watched and the resources available. Baluran National Park spans savanna, monsoon forest, evergreen forest, and mangrove, with banteng (Bos javanicus) as its keystone species. Conventional survey methods are labour-intensive, limited in reach, and hard to repeat consistently over time.

Unmanned aircraft and satellite imagery open the possibility of repeated monitoring at far lower cost. The volume of data produced, however, exceeds what manual interpretation can handle, which makes machine-learning automation a necessity rather than a preference.

1.2 Objectives

1. Design and implement an integrated multi-sensor detection pipeline covering three pillars: per-species fauna, vegetation and tree crowns, and monitoring for indications of illegal logging.
2. Test the pipeline's feasibility on real Indonesian field data rather than simulated data.
3. Document every methodological decision together with its reasoning and the alternatives rejected, so that others can scrutinise and reproduce it.
4. State the system's limits of authority explicitly: what may and may not be concluded from its output.

1.3 Statement of contribution

The contribution of this work is APPLIED SYSTEMS ENGINEERING, not algorithmic novelty. Every model used (YOLO11, Mask R-CNN via Detectree2, BioCLIP 2, Random Forest) is an established method. What is new is:

- Integration of the three pillars into a single operational platform with a preserved chain of custody (SHA-256) from the moment data enters.
- Application to the Indonesian tropical savanna domain, which is not represented in the pre-training data of those models.
- A cause-and-effect triage flow applied before any indication of violation is recorded, which arose from a real pilot finding and prevents false accusations.
- Quantitative documentation of limitations, including negative findings and outright failures.

CAUTION: This work does NOT claim algorithmic novelty, does NOT claim performance comparable to international publications, and does NOT claim generalisation beyond the test data.

1.4 Scope

The case study is confined to Baluran National Park, with emphasis on the Bekol savanna. Fieldwork at Baluran was closed before this document was written, so the available data is fixed and cannot be extended for the purposes of this study.

2. Literature Review and Positioning

2.1 Individual tree crown delineation

Systematic reviews of deep-learning tree crown detection show instance segmentation as the dominant approach over 2020-2025, with RGB imagery the primary input in most studies. Detectree2 (Ball et al., 2023) is a Mask R-CNN implementation trained specifically for tropical forest and is used here as the pre-trained starting point.

Ball et al. (2023) report performance across four tropical forest sites:

Site	Test trees	Precision	Recall	F1
Paracou (French Guiana)	381	0.595	0.543	0.568
Danum (Malaysian Borneo)	278	0.713	0.662	0.687
Sepilok East	167	0.612	0.653	0.632
Sepilok West	704	0.640	0.656	0.648
Mean	(1,530)	0.640	0.629	0.634

That study used 3,797 manually delineated crowns for training, crewed aerial imagery at 8-10 cm GSD, and a Tesla P100 GPU with training time under two hours.

CAUTION: The F1 figures above are NOT equivalent to the AP50 figures in Chapter 6. AP50 averages precision along a curve at an IoU threshold of 0.50, whereas F1 is computed at a single confidence threshold. The two must not be placed side by side as a direct comparison. The comparisons that ARE legitimate are discussed in Chapter 6.

2.2 Wildlife detection from unmanned aircraft

This field developed actively over 2024-2025, including comparisons of recent YOLO variants on nadir savanna imagery. Performance figures across studies are not directly comparable, however, because target species, flight altitude, spatial resolution, and class definitions differ. The fauna literature is therefore cited as qualitative context, not as a quantitative benchmark.

2.3 Biological foundation models

BioCLIP 2 is a vision-language model trained with hierarchical taxonomic supervision on a large-scale collection of organism imagery, designed for few-shot and zero-shot classification. It is used here as a second-stage classifier following generic detection.

3. Study Area and Data Sources

3.1 Study area

Baluran National Park, East Java, Indonesia. Analysis focuses on the Bekol savanna, the principal banteng habitat and the area with the most complete data coverage.

3.2 Data inventory

Asset	Sensor	Count	Notes
Baluran NP data (2 Nov 2023)	RGB + NIR + thermal	139 frames/sensor	Co-registration verified via GPS EXIF; the only thermal session
Block 286_1	RGB + NIR	403 + 403	Co-registration complete
Block 286_2	RGB + NIR	403 + 403	Co-registration complete
Block 286_3	RGB + NIR	397 + 397	Co-registration complete
Block 286_4	RGB + NIR	397 + 378	19 NIR frames missing
Block 268_1	RGB + NIR	401 + 251	NOT co-registered - rejected for fusion

3.3 Verification of sensor co-registration

The claim that two sensors came from a single flight was NOT taken from file naming but verified against per-frame GPS EXIF metadata: frame counts and coordinates were compared across sensors. This decision followed the discovery that timestamps on several devices were not synchronised, making date labels potentially misleading.

Note: For areas flown in a grid pattern, comparing only the first and last points is inadequate; the full bounding box of all frames is required. This mistake occurred at an early stage and has been corrected.

3.4 Inclusion criteria and rejected data

Data was included when it met two conditions: valid georeferencing, and spatial resolution adequate for the role assigned to it. The resolution threshold for primary wildlife detection was derived from adult banteng body dimensions (0.6-0.9 m across) with a target of roughly 15 pixels across the object.

Data rejected for a given role was recorded, not discarded:

- Block 268_1 was rejected for multispectral fusion because RGB and NIR frame counts and coordinates did not match.
- raster_baluran.tif was rejected as a fusion source because it has no co-registered NIR/thermal counterpart; the nearest thermal GPS point lies roughly 167 m outside its bounds.
- Very coarse assets (around 42 cm/pixel) were rejected for detection and retained as context.

3.5 Spatial resolution uncertainty

CAUTION: Every GSD value in this document is INDICATIVE. EXIF altitude is referenced to mean sea level, not to height above ground, which is what actually determines GSD. The difference depends on terrain elevation, which has not been verified against an elevation model. GSD figures must not be cited as exact measured values.

Asset	Indicative GSD	Basis
rgb-001.tif	1.54 cm/px	Measured directly from the geotransform
raster_baluran.tif	3 cm/px	Measured directly from the geotransform
Block 286_1 RGB	~4.8 cm/px	Computed from manufacturer specification + EXIF altitude
Block 286_1 NIR	~7.3 cm/px	Computed from the manufacturer's stated HFOV
Thermal	~9 cm/px (6.4-11.9)	Geometric estimate; sensor FOV absent from EXIF

3.6 Annotation data

Item	Count	Notes
Total consolidated annotations	722	A single COCO file with a per-annotation audit trail
- fauna	145	5 source images
- flora	577	3 source images
Excluded	139	Source imagery not found

Annotations without source imagery were excluded because they can serve neither training nor evaluation, even though the annotations themselves are intact.

4. Methodology by Pillar

4.1 Common architecture

All pillars share one processing pattern: data enters with a cryptographic fingerprint, is pre-processed for scale consistency, is passed through a model or spectral index, is merged and georeferenced, and is then presented as a confidence-bearing HYPOTHESIS awaiting human verification. No path bypasses the human verification stage.

Pillars differ in data source, model, and output geometry (points for fauna, polygons for vegetation, change areas for cover monitoring).

4.2 Fauna pillar: two-stage architecture

Fauna detection is split into two stages: a class-agnostic detector that finds the presence of an animal, and a separate classifier that determines species identity from the image crop. This separation was chosen because adding a new species then requires only updating the classifier's data, without retraining the detector or re-annotating existing boxes.

Alternative	Characteristics	Decision
Single multi-class detector	One model, one training cycle	Rejected - adding a species demands full re-annotation and retraining
Generic detector + separate classifier	Two models, independent update cycles	Chosen - species expansion is cheap and compatible with existing generic annotations

4.3 Resolution-aware tiling

Gigapixel rasters cannot be fed to a network whole, so they are cut into tiles using overlapping windows. Tile size is fixed against GROUND DISTANCE, not pixel count.

This is not an aesthetic choice but the consequence of a measured finding: with tiles fixed at 800 pixels, a raster at 5.50 cm/pixel yielded 45 detections while a raster at 1.54 cm/pixel yielded 297 detections from the same model. The difference does not reflect a change in model capability but the apparent size at which objects appear. If tiles remain fixed in pixels, detection quality silently depends on whatever resolution a user happens to upload.

CAUTION: Scale normalisation does NOT restore detail lost when resolution is reduced. Its purpose is to make the pipeline CONSISTENT across resolutions, not to increase resolution.

4.4 Cross-tile merging and spatial deduplication

Boxes from all tiles are merged with global Non-Maximum Suppression: boxes overlapping above an IoU threshold within the intersection between tiles are treated as the same object, and only the box with the highest confidence is kept.

For overlapping loose-photo transects, one individual may be photographed repeatedly. Detections across frames are merged by single-linkage clustering: two points join the same cluster when a chain of points exists in which each step falls below a given radius, distances computed with the haversine formula on the Earth's surface. Points without GPS coordinates are never force-merged.

4.5 Vegetation pillar: crown segmentation

Crown segmentation uses Mask R-CNN via Detectree2 with tropical-forest pre-trained weights, subsequently fine-tuned on Baluran flora annotations. Its output is crown polygons, not merely boxes.

Two adjustments are mandatory before tiling:

1. Reprojection to a projected coordinate system. The tiling tool interprets tile size directly as units of the source coordinate system; on a raster in a geographic system, a tile of size 100 would mean 100 DEGREES rather than 100 metres. The UTM zone is determined automatically from the data's position rather than fixed to one zone, so the pipeline stays correct when applied to other areas.
2. Resolution normalisation. Rasters far finer than the reference scale are downsampled within the same reprojection pass. Without this, a 100 x 100 metre tile on a 1.54 cm/pixel raster becomes roughly 6,500 x 6,500 pixels and the process halts on memory exhaustion - which actually occurred in a production run. After normalisation the tile becomes roughly 2,083 x 2,083 pixels.

4.6 Improving crown annotation quality

The available flora annotations came from a format storing only bounding boxes, so the masks the model learned were rectangles rather than true crown silhouettes. To assess the impact, masks were upgraded to silhouettes using GrabCut initialised from the existing boxes.

GrabCut was chosen not because it is the most modern method but because its process can be explained and it requires no additional large model - consistent with the principle of preferring explainable approaches over opaque ones.

Configuration	Masks becoming silhouettes	Remaining rectangular
Fixed 6-pixel context margin	244 of 577 (42.3%)	333
Proportional margin (35% of shortest side)	305 of 577 (52.9%)	272

The move from a fixed to a proportional margin came from quantitative diagnosis, not trial-and-error tuning: in the initial configuration, 333 of 577 annotations produced EMPTY masks with a median ratio of 0.0. The cause was that with a fixed margin, small boxes nearly filled the sub-image, leaving the algorithm no background samples from which to build its colour model. A proportional margin guarantees a background ring at any box size.

Note: Unconvincing annotations were NOT forced into polygons but returned as rectangles and flagged explicitly, so that the training stage can filter them or weight them differently.

4.7 Cover monitoring and illegal logging indication

This pillar is not a standalone detector but a derivative of the vegetation pillar, worked through change detection over time on medium-resolution satellite imagery, followed by staged verification.

Two normalised spectral indices are used: NDVI as a marker of greenness and canopy loss, and NBR as an indication of burnt area. Area-mean values are computed on two dates and their difference compared against a threshold.

4.8 Cause-and-effect triage before recording an indication

Every change alert passes through automatic, auditable rule-based filtering BEFORE it is called a candidate violation:

- Geometric intersection against permit or concession layers. Where they intersect, the change falls within licensed activity and is not an indication of violation.
- Seasonal calendar test. An index decline coinciding with the wet-to-dry transition is flagged as possible seasonal drying.
- The remainder is forwarded as a candidate, triggering aerial and then ground verification.

The seasonal rule was not designed up front but emerged from a real pilot finding: the NDVI of the Baluran savanna fell from 0.394 in April to 0.213 in August. Without this rule that decline would have been classified as an indication of illegal logging, when it is purely a dry-season phenomenon in savanna.

CAUTION: Method limits: medium-resolution satellite imagery can capture change only at scales of roughly 0.1 hectare and above. Selective felling of one or two trees will NOT be visible. Furthermore, determining legal status is NOT a system output - the system documents, authorised humans decide.

4.9 Human-in-the-loop annotation flow

Training-data annotation runs as a permanent flow rather than a temporary first step: the model produces candidate annotations first, and humans act as reviewers who correct, not as draughtsmen starting from nothing. Human attention is allocated by confidence - low-confidence and anomalous predictions are reviewed first.

The basis for this decision is a human limitation both widely documented in the literature and observed directly in this project: annotation consistency declines after a few dozen objects as reviewer fatigue sets in.

4.10 Ecological statistical methods

Species inventory completeness is estimated using the Chao2 estimator, the INCIDENCE-based variant appropriate to the structure of survey-transect data. The better-known Chao1 estimator is abundance-based and is NOT appropriate for this data structure - a distinction often confused.

S_Chao2 = S_obs + Q1 (Q1 - 1) / (2 (Q2 + 1)), where S_obs is the number of observed species, Q1 the number appearing exactly once, and Q2 exactly twice. Completeness is computed as S_obs divided by S_Chao2. The implementation was checked against manual calculation, including the edge case where Q2 equals zero.

5. Evaluation Protocol

This chapter precedes the results because performance figures are meaningless without the protocol that produced them. Every figure in Chapter 6 was computed with the protocol below.

5.1 Data splitting

The train/validation split is made at the level of SOURCE IMAGE, not tile. The reason: tiles are produced with overlapping windows, so a per-tile split would place the same image content on both sides and leak data.

Pillar	Training	Validation
Fauna	4 source images - 92 tiles (60 empty), 155 boxes	1 source image - 23 tiles (15 empty), 44 boxes
Vegetation	2 source images - 70 tiles, 573 boxes	1 source image - 35 tiles, 244 boxes

CAUTION: There is NO independent test set. Training and validation data come from the same photographic session, on the same day, at adjacent locations. This evaluation is IN-SAMPLE with a weak split; the figures measure model capability under those imaging conditions, NOT the capacity to generalise to other field conditions.

5.2 Detection metric protocol

Detection metrics depend on protocol, so the protocol is stated explicitly to permit reproduction:

- Confidence threshold 0.30 - identical to the production pipeline threshold AT THE TIME OF MEASUREMENT (YOLO11n). The production detector was later replaced and uses a different threshold; see section 6.7.
- Matching criterion IoU >= 0.50 against ground-truth boxes.
- Greedy matching, following standard COCO evaluation: predictions are processed from highest to lowest score, each prediction immediately claims the best-matching ground-truth box available at that moment, and a ground-truth box may be matched only once. This approach is chosen NOT because it is optimal - global optimal assignment such as the Hungarian algorithm can produce a better overall pairing - but because it is what standard COCO evaluation does, which keeps the resulting figures comparable with the literature.
- Unmatched predictions count as false positives; unmatched ground-truth boxes count as false negatives.

Note: Figures in Chapter 6 were recomputed with the protocol above. Values reported by the training library's built-in tooling may differ because they use precision-recall curves and different thresholds. Both approaches are valid; what is not defensible is mixing them without stating the protocol.

5.3 Confidence intervals

Every metric is reported with a 95% confidence interval computed by non-parametric bootstrap: B = 5,000 replicates, resampling at tile level, fixed seed 20260818 for repeatability.

CAUTION: Tile-level resampling measures uncertainty only WITHIN that validation image. For generalisation the independent unit is the image or flight session, and the validation set contains only one source image (n = 1). A GENERALISATION confidence interval therefore cannot be computed from the available data. This is a limitation of data collection design, not a shortcoming of the statistical analysis.

5.4 Segmentation metrics

For the vegetation pillar, Average Precision at an IoU threshold of 0.50 (AP50) is used for both boxes and masks, following COCO convention. Evaluation runs every 100 iterations on the validation set, and the checkpoint reported is the one with the best validation performance, not the last checkpoint. This choice matters because an overfitting trend was detected explicitly in the vegetation pillar (Chapter 6).

6. Results

CAUTION: This chapter contains ONLY figures that are legitimately measurable. Several commonly reported metrics are deliberately OMITTED because inspection found the measuring instrument invalid for this data. Every omission is stated openly with its reason and elaborated in Chapter 7. The absence of a figure here is not an oversight but a refusal to report a number that does not measure what it appears to measure.

6.1 Fauna detection

Fauna validation annotations were checked for completeness before any metric was reported. Across all detections from both models, not one animal was detected with high confidence yet lacked an annotation: zero for model v2, and only one for model v1, at low confidence. No false positive exceeded 0.70 confidence. The false positives present are duplicate detections and boxes that missed on animals that were in fact annotated. All metrics in this section are therefore legitimate, including Precision.

The protocol must be stated because metrics change with protocol: confidence threshold 0.30 matching the production pipeline at the time of measurement (YOLO11n; see 6.7), matching criterion IoU >= 0.50, greedy matching by highest score, bootstrap of 5,000 replicates at tile level, random seed 20260818.

Model	TP	FP	FN	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)
YOLO11n v1	34	6	10	0.850 [0.786-0.917]	0.773 [0.697-0.902]	0.810 [0.742-0.905]
YOLO11n v2	39	2	5	0.951 [0.880-1.000]	0.886 [0.838-1.000]	0.918 [0.869-1.000]

Ground-truth boxes number 44, spread across 8 of 23 validation tiles.

The confidence intervals of the two models overlap on all three metrics. Concluding 'no difference' from that overlap would be wrong, however, because it ignores that both models are evaluated on the SAME TILES, so their errors are correlated. The difference between versions is therefore computed as an interval on the DIFFERENCE itself, using resampling indices shared by both models.

Metric	Difference (v2 minus v1)	95% CI	Conclusion
Precision	+0.101	[+0.044, +0.190]	Significant
Recall	+0.114	[+0.036, +0.200]	Significant
F1	+0.108	[+0.057, +0.187]	Significant

Note: No difference interval contains zero. The v1 to v2 improvement in the fauna pillar can therefore be stated as statistically significant.

CAUTION: The limits of interpretation still apply in full. Significant here means the difference is real ON THE IMAGE TESTED, not that the improvement will recur at another location, season, or with other equipment. The v2 interval touching 1.000 in the preceding table is a classic sign that the test data is too small to distinguish the model from perfect, NOT a sign that the model is perfect.

6.2 Tree crown segmentation

Model v2 was trained for the full 600 iterations without early stopping, taking 3 hours and 53 minutes on a general-purpose processor without an accelerator. Its only difference from v1 is the shape of the training masks: v1 used rectangular masks derived from bounding boxes, v2 used crown silhouettes. All other hyperparameters were held identical, and it was verified that the 573 training boxes and 244 validation boxes are identical between the two versions, so that any difference observed genuinely originates in mask shape.

The comparison evaluates both models against the SAME ground truth, namely 144 genuine silhouette polygons. This differs from earlier reporting, in which v1 was evaluated against rectangular masks and v2 against silhouettes, leaving the figures never comparable. Matching uses mask IoU rather than box IoU, because shape is what is under test. Differences are computed pairwise on identically resampled tiles.

Metric	v1 (rectangle-trained)	v2 (silhouette-trained)	Paired difference (95% CI)
Recall	0.771	0.861	+0.090 [+0.045, +0.140]
Confidence of correct detections (median)	0.775	0.854	not formally tested
Correct detections above 0.90 confidence	8	38	not formally tested

Note: The recall difference interval does NOT contain zero. This is the only between-version comparison in the vegetation pillar that can be stated as statistically significant.

What may legitimately be concluded: training with silhouette masks makes the model find significantly more of the annotated crowns, and makes it more confident on the crowns it gets right. Nothing beyond that.

CAUTION: Precision, F1, and all Average Precision comparisons for the vegetation pillar are deliberately NOT reported. Inspection found the crown validation annotations incomplete: two-thirds of the detections counted as false positives are in fact real tree crowns that were never annotated, confirmed by visual inspection. Metrics containing a false-positive term therefore measure annotation completeness more than model capability. Recall is unaffected because it depends only on whether ANNOTATED objects were found. Full discussion in Chapter 7.

6.3 Fauna species classification

The exemplar-based approach outperformed the text-description approach on this data. The underlying data is very small, however, so the figures must NOT be cited without their sample size: only three positive examples exist, of which two were correct and one was missed.

CAUTION: The Precision value of 1.000 for the exemplar-based approach derives from TWO correct cases. That number is meaningless as a measure of capability and must not be quoted apart from its sample size.

A three-class variant adding Javan rusa was REJECTED, not because its figures were poor but because a domain confound was found: exemplars drawn from a different raster inverted the prediction distribution wholesale, on thin confidence margins. Accepting such a result would mean reporting an artefact of data collection as model capability.

6.4 Satellite-based cover monitoring

A real pilot over the study area savanna showed the greenness index falling from 0.394 in April to 0.213 in August. A decline of that magnitude is fully explained by the dry season.

CAUTION: This is an important NEGATIVE finding, and precisely there lies its value: without a seasonal filter rule, the system would allege logging on a decline that is purely seasonal. Any deployment of multi-temporal monitoring in a savanna landscape with pronounced seasons must include such a filter.

6.5 Operating threshold for production

The confidence threshold used by the vegetation production pipeline was inherited from the fauna pipeline and never tuned for crowns. A threshold sweep was performed reporting ONLY recall and the number of candidates a human must review. Precision and F1 were deliberately not used as the basis for selection: because the annotations are incomplete, choosing the threshold that maximises F1 would tune the model to fit defective annotations, penalising it precisely for finding trees that annotation missed.

Threshold	Recall	Crowns found	Candidates to review	Against current threshold
0.30 (current)	0.861	124 of 144	626	100%
0.40	0.847	122	485	77%
0.50	0.833	120	395	63%

Threshold	Recall	Crowns found	Candidates to review	Against current threshold
0.60	0.799	115	307	49%
0.70	0.694	100	225	36%
0.80	0.590	85	134	21%

The knee lies between 0.50 and 0.60. Raising the threshold to 0.50 costs 3.3 per cent of recall while cutting the number of candidates by 37 per cent.

CAUTION: The 'candidates to review' column must NOT be read as wasted effort. Because two-thirds of the detections counted as false positives are real unannotated crowns, raising the threshold cuts verification workload AND cuts genuine tree discoveries at the same time. How much is discarded cannot be measured with the annotations available. Threshold selection is therefore a question of the managing institution's review capacity, not a statistical question, and this document does not set it.

6.6 Computational performance

All production processing runs on a general-purpose processor without a graphics accelerator. One crown segmentation job over a 6.24 gigabyte raster completed in 6 minutes 31 seconds with peak memory use of 1,235 megabytes. Before resolution normalisation was applied, the same job exhausted memory and killed the worker process.

These figures are included because they bear on deployment feasibility for area-management institutions, which generally do not possess specialised hardware.

6.7 Replacement of the production fauna detector

The model-library licence audit (5 September 2026) found that the package running YOLO11n is licensed under AGPL-3.0. On the package owner's own reading, serving users over a network already triggers the obligation to release the full source of the derivative work, including model weights trained with it. The trigger is network use, not sale. This section records the licence facts and the owner's official reading; it is NOT legal advice.

Alternative	Consequence	Decision
Buy the enterprise licence	Recurring cost; dependence on a single vendor remains	Rejected
Open the entire platform code	Touches data sovereignty and access control (Chapter 8)	Rejected
Continue while aware of the exposure	Risk stays attached to the fauna pillar and its weights	Rejected
Replace the detector with a permissively licensed architecture and retrain	All fauna figures must be re-measured under a new protocol	Chosen - first tested in a pre-registered detector comparison

The comparison rules were written and locked BEFORE the data were seen (pre-registration): three leave-one-image-out folds with 137 test animals, pooled AP50 as the primary metric (101-point interpolation), a non-inferiority margin of -0.05 AP50 against the reference, three random seeds (42, 43, 44), pseudo-labels excluded from training, and sequential gates: licence, ONNX export fidelity, non-inferiority, then speed and memory on a general-purpose processor. Entrants: A0 YOLO11n (reference), K1 RF-DETR Nano, K2 MIT-licensed YOLOv9-T, and K3 D-FINE-N. K2 was eliminated in the seed-42 round; by rule, at most two finalists advance.

Model	Seed 42	Seed 43	Seed 44	Range across seeds
A0 YOLO11n (reference)	0.6804	0.7047	0.5833	0.5833-0.7047
K1 RF-DETR Nano	0.7019	0.7042	0.7041	0.7019-0.7042
K3 D-FINE-N	0.7202	0.6775	0.7106	0.6775-0.7202

Gate	K1 RF-DETR Nano	K3 D-FINE-N
1. Code licence	Apache-2.0 - pass	Apache-2.0 - pass
2. ONNX fidelity (AP50 difference <= 0.01)	0.0000 - pass	0.0000 - pass
3. Non-inferiority (95% CI, margin -0.05)	+0.0473 [-0.0079, +0.1068] - pass	+0.0466 [-0.0022, +0.1037] - pass
4a. Time per tile <= 3 times reference (reference 66.4 ms)	597.0 ms (9.0 times) - FAIL	121.3 ms (1.83 times) - pass
4b. Peak memory <= 2 GB	920.6 MB - pass	502.8 MB - pass

No candidate was shown to be BETTER than the reference: the lower bound of both intervals is still below zero. What was shown is non-inferiority within the stated margin only. The deciding gate was speed: K1 is nine times slower than the reference on a general-purpose processor. D-FINE-N passed every gate and was designated the production fauna detector (18 September 2026). This agrees with conclusion 9.1 item 2: differences between architectures are small compared with the effect of data.

The production model was retrained on ALL human-labelled data (115 tiles containing animals, 75 empty tiles, 199 boxes, five source images), exported to ONNX, and run with ONNX Runtime 1.30.0 on a general-purpose processor. The production code was tested against the experiment wrapper on 60 real tiles: 113 detections against 113 detections, largest score difference 0.0001. The AGPL-3.0 package was uninstalled from the production environment on 20 September 2026; the YOLO11n weights are kept only as a scientific comparator and can no longer be loaded by the application.

Confidence scores from the two architectures are NOT comparable, so YOLO11n's 0.30 threshold was not reused. The new threshold was chosen from out-of-fold predictions (137 animals, mean of three seeds):

Model and threshold	Precision	Recall	F1
A0 YOLO11n at 0.30 (former production threshold)	0.696	0.764	0.728
K3 D-FINE-N at 0.30	0.607	0.837	0.704
K3 D-FINE-N at 0.50 (CHOSEN)	0.692	0.830	0.754
K3 D-FINE-N at 0.70 (highest F1)	0.777	0.788	0.783

The platform manager chose 0.50 because it gives precision equal to the former detector (0.692 against 0.696) with higher recall (0.830 against 0.764). Every detection is still reviewed by a human, so a missed animal costs more than a surplus candidate. A threshold of 0.70 maximises F1 but gives up about four in every hundred animals that are still caught at 0.50.

CAUTION: Three limits on interpretation. FIRST, the production model has NO test figures of its own: all labelled data were used to train it, so the figures in this section come from the fold models and are an ESTIMATE of performance, not a measurement of the model that is running. SECOND, the figures in 6.1 (44 boxes, one source image, pseudo-labels used to train v2) and the figures here (137 animals, three folds) follow different protocols and must NOT be compared directly; the drop in YOLO11n's figures from 6.1 to this section reflects a stricter protocol, not a degraded model. THIRD, the test set remains small and comes from the same flight sessions, so limitation 7.1 applies in full.

7. Limitations

This chapter is written on the assumption that readers will look for weaknesses in this work. Stating them first, complete with magnitude and consequence, is more useful than waiting for others to find them. Several were discovered at a late stage and invalidate figures previously reported; that is stated as it happened.

7.1 No independent test set

Training and test data come from the same photographic session. The fauna validation set comes from one image, as does the vegetation validation set. Every figure in Chapter 6 therefore describes system capability ON THAT IMAGERY, not the capacity to generalise to other locations, seasons, flight altitudes, or equipment.

CAUTION: A GENERALISATION confidence interval cannot be computed at all from this data. The independent unit for generalisation is the image or flight session, and only one is available. The tile-level resampling used in Chapter 6 measures uncertainty only WITHIN that image. This is a limitation of data collection design, not of the analysis, and no statistical method can repair it.

7.2 Crown validation annotations are incomplete

This is the most consequential limitation in this document, and it was discovered only after all model training had finished. Inspection of detections counted as false positives found two-thirds of them falling in areas with no annotation at all. Visual inspection of those tiles confirmed that most are real tree crowns that were never annotated. Annotation appears biased towards large crowns, while small and medium crowns were frequently missed.

Consequently, every metric containing a false-positive term for the vegetation pillar measures annotation completeness more than it measures model capability. Precision, F1, and Average Precision comparisons were therefore omitted from Chapter 6.

The same inspection was applied to fauna annotations and produced a different result: the fauna annotations are COMPLETE, so fauna metrics remain legitimate. The difference makes visual sense. Animals appear at high contrast against the background and can be counted one by one, whereas tree crowns in savanna overlap one another, have ambiguous boundaries, and are far more numerous per tile, making them far more prone to being missed.

Note: What resolves this limitation: exhaustively annotating a single tile, every crown without exception. That makes Precision legitimate and lets the production threshold be chosen from figures. Estimated at a few hours of one annotator's time.

7.3 Vegetation ground truth is not uniform

Crown silhouettes were obtained by upgrading legacy box annotations using a classical segmentation method. That method succeeded on 52 per cent of training annotations and 59 per cent of validation annotations; the remainder fell back to rectangles. Silhouette annotations cover about 61 per cent of their box area, whereas the failures cover 100 per cent.

CAUTION: For visually similar objects of the same kind, the model therefore receives two contradictory lessons: some crowns are taught that their shape fills about 61 per cent of the box, others that it fills the whole box. That difference is not a property of the crowns but a defect of the annotation process, and it is the strongest candidate explanation for the unusual behaviour of the segmentation metrics.

A second consequence must be stated: the segmentation Average Precision reported by the software by default counts all 244 objects, including the 100 still rectangular. Measurement restricted to the 144 silhouette objects yields substantially higher figures. A gap that large means the default figure systematically understates segmentation capability, and figures of that kind in earlier reports must not be used as a measure of capability.

7.4 Metrics deliberately not reported

The following list is included so that readers can confirm the omissions are deliberate:

Metric	Pillar	Reason for omission
Precision, F1	Vegetation	Annotations incomplete; both measure annotation completeness
AP comparison v1 against v2	Vegetation	Ground truth differs in shape between versions; figures not comparable
Confidence intervals for AP	Vegetation	Not yet computed
Any metric	Thermal	No model exists at all
Any metric	Crown species classification	Never trained; zero field labels
Generalisation confidence interval	All pillars	Only one source image available

7.5 A correction to the testing method, and the conclusion it changed

In the original drafting, between-version comparison in the fauna pillar was judged by checking whether two separate confidence intervals overlapped, while the vegetation pillar used a paired difference interval. The two pillars were therefore not tested in an equivalent way.

The paired approach is stronger. Checking the overlap of two separate intervals ignores that both models are evaluated on the same tiles, so their errors are correlated; it is exactly that correlation which narrows the difference interval. As a result two intervals may overlap even when the paired difference is real.

CAUTION: This imbalance HAS BEEN CORRECTED, and the correction CHANGED the conclusion. Recomputation with paired differences found the fauna v1 to v2 improvement SIGNIFICANT on all three metrics, whereas the earlier approach concluded it was not significant. The earlier statement is withdrawn. It is recorded here rather than concealed because it demonstrates that the statistical conclusions of this work are sensitive to the choice of test - and readers are entitled to know that.

Both pillars are now tested equivalently. No per-model figure changed; only the way differences between versions are judged.

7.6 Species identities not confirmed by experts

Not one species identity in this document has been confirmed by a taxonomist or field officer. Every species mention has the status of an indication based on visual similarity from the air. Distinguishing species of similar size and colour in aerial imagery remains an open problem that this project has not solved.

7.7 Pillars without a model

- The thermal pillar has no model at all. Thermal data exists from only one flight session because the equipment failed afterwards, and a separate study concluded that airborne thermal is severely limited beneath dense canopy.
- Crown species classification has never been trained. Its features are extracted, but field survey species labels number zero, and the program deliberately refuses to run until a minimum number of labels exists.
- Camera-trap wildlife detection is at study stage only; the equipment has not been procured.

7.8 Physical measurement limitations

- All ground sample distance values are indicative because the difference between height above mean sea level and height above ground has not been resolved.
- Crown height is computed from a raw digital surface model, NOT a canopy height model, because no digital terrain model is available. The values therefore contain a ground-elevation component and must not be read as true tree height.

- Some crown polygons in the production run are implausibly large for a single tree, indicating that clump merging remains unhandled.

7.9 Operating threshold not established

As set out in section 6.5, the production confidence threshold for the vegetation pillar cannot be chosen from the available data because that choice depends on the managing institution's verification capacity. This document presents the trade-off quantitatively but does NOT set the value.

7.10 Summary: what resolves what

Limitation	What resolves it	Estimated effort
Crown annotations incomplete	Exhaustive annotation of a single tile	A few hours
Ground truth not uniform	Genuine crown polygon annotation rather than box-derived	A few days
No independent test set	A second flight session at a different place or time	One field activity
Species identities unconfirmed	Verification by experts or park officers	Depends on expert availability
Pillars tested unequally	COMPLETED 5 September 2026 - fauna recomputed with paired differences; the conclusion changed	-
Crown species classifier untrained	Dendrological survey for field labels	One field activity
Thermal pillar without a model	Camera-trap procurement as a complement	Procurement and installation
Production fauna model has no test figures of its own (6.7)	A test set from another flight session, never used for training	One field campaign

8. Data Governance, Ethics, and Limits of Authority

8.1 Limits of system authority

This section is decisive and must not be abridged when quoted.

CAUTION: This system produces measured HYPOTHESES, not determinations. Its output must NOT serve as the sole basis for: determining legal status or violation, enforcement action against persons or businesses, determining area boundaries or land rights, or scientific claims about species populations. For all such purposes the system's output has the standing of an initial indication that must be verified by authorised humans.

This formulation is not mere legal formality but a direct consequence of measurable technical limitations: not one species identity in this system has been confirmed by an expert, and all performance figures come from small validation sets.

8.2 Chain of custody

Every incoming file has its cryptographic fingerprint computed once using SHA-256, and that value is never altered afterwards. A change of even one bit alters the fingerprint, so the authenticity of the material can be demonstrated later.

The chain of custody is built from the ingestion stage rather than added afterwards, because evidence assembled after the fact loses its evidentiary value.

8.3 Protection of sensitive species locations

Precise coordinates of species that are threatened or of high economic value are not shown to users outside the area management group. Coordinates are generalised to grid cell centres, and every access to precise coordinates is written to an audit log.

This protection applies to fauna and to timber-valuable flora alike, because disclosing locations can facilitate poaching or felling - a real risk inherent in publishing biodiversity data.

8.4 Data sovereignty and ownership

All raw data, trained models, and result databases reside on infrastructure controlled by the operator within Indonesian territory. Third-party services used are limited to open sources for satellite imagery and global biodiversity databases; no field data belonging to the protected area is transmitted to those services.

8.5 Access control

Access is tiered with separation by managed area. The ability to modify taxonomy, review annotations, access cross-area data, and manage users are separated across different levels of authority, so that an error at one level does not affect the system as a whole.

8.6 Openness and repeatability

Every methodological decision, including those later revised or rejected, is documented with its reasoning. Software versions, random seeds, hardware specifications, and data file fingerprints are listed in the appendix so that others can assess or repeat the procedures carried out.

Note: Negative findings and failures are reported on equal footing with successes. The rejected three-class few-shot classification experiment contaminated by domain difference, the process failure from memory exhaustion, and the empty masks in the initial annotation upgrade configuration are all recorded in this document.

9. Conclusions and Further Work

9.1 Conclusions

This work shows that an integrated multi-sensor detection pipeline for conservation monitoring can be built and operated on real Indonesian field data, with limited computational resources and without a graphics accelerator in the production environment.

What was demonstrated:

1. All three pillars run end to end from data ingestion to georeferenced output, with the chain of custody preserved from the moment a file enters.
2. Fine-tuning pre-trained models on local data yields performance gains far exceeding the differences between model architectures, directing resource priority towards data quality and quantity rather than the search for new architectures.
3. Cause-and-effect triage prevented a false indication of violation in a real case of seasonal vegetation index decline.
4. Spatial scale awareness at the tiling stage proved decisive; without it, detection quality silently depends on the resolution of whatever file happens to be uploaded.

What could not be demonstrated is stated with equal openness in Chapter 7, and is not claimed as success.

9.2 Further work by priority

Priority	Step	Prerequisite
High	Build an independent test set from a different flight session	New data acquisition in another area
High	Verification of species identity by experts or park officers	Cooperation with the managing authority
High	Human crown polygon annotation to replace the automated approach	Trained annotator time
Medium	Training the crown species classifier	At least 30 field dendrology labels
Medium	Thermal detection model	A new thermal acquisition with a radiometric sensor
Medium	True crown height computation	A digital terrain model
Low	Object tracking in video	Video data with telemetry

9.3 Closing statement

The system described in this document is and will remain under development. The author holds that the usefulness of a conservation monitoring system is determined not by the size of the figures it reports but by how far the limits of its reliability are known and stated, so that users can judge how far its output may be trusted.

10. References

Note: References marked [directly verified] were read from the original text during the preparation of this document. The remainder are standard works well known in their fields; page and issue details MUST be checked again against the original before this document is submitted to a publisher.

- Ball, J. G. C., Hickman, S. H. M., Jackson, T. D., Koay, X. J., Hirst, J., Jay, W., Archer, M., Aubry-Kientz, M., Vincent, G., & Coomes, D. A. (2023). Accurate delineation of individual tree crowns in tropical forests from aerial RGB imagery using Mask R-CNN. Remote Sensing in Ecology and Conservation. https://doi.org/10.1002/rse2.332 [directly verified]
- He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV).
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV).
- Rother, C., Kolmogorov, V., & Blake, A. (2004). GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 23(3).
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
- Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics, 43(4), 783-791.
- Stevens, S., Wu, J., Thompson, M. J., Campolongo, E. G., Song, C. H., Carlyn, D. E., et al. (2024). BioCLIP: A vision foundation model for the tree of life. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Key, C. H., & Benson, N. C. (2006). Landscape assessment: Sampling and analysis methods. USDA Forest Service General Technical Report RMRS-GTR-164-CD.
- Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. NASA Special Publication, 351.
- Jocher, G., & Qiu, J. (2024). Ultralytics YOLO11 [software]. https://github.com/ultralytics/ultralytics
- Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., & Wu, F. (2024). D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. arXiv:2410.13842 (ICLR 2025).
- Robinson, I., Robicheaux, P., & Popov, M. (2025). RF-DETR [software]. https://github.com/roboflow/rf-detr
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2 [software]. https://github.com/facebookresearch/detectron2

11. Reproducibility Appendix

11.1 Software environment

The exact versions in use when every figure in this document was produced:

Component	Version
Python	3.12.1
PyTorch	2.13.0+cpu
torchvision	0.28.0+cpu
Detectron2	0.6
Detectree2	2.1.2
Ultralytics	8.4.117 - only for the figures in 6.1 and the reference in 6.7; uninstalled from production 20 September 2026
ONNX Runtime	1.30.0 - runs the production fauna detector (D-FINE-N), CPU provider
pybioclip	2.1.5
rasterio	1.5.1
OpenCV	5.0.0.93
scikit-learn	1.9.0
pycocotools	2.0.11
geopandas / shapely	1.1.4 / 2.1.2

11.2 Hardware environment

Environment	Specification	Role
Workstation	16-core processor, 32.6 GB RAM, Windows 11; legacy GPU without tensor cores	Model training and photogrammetry processing
Production server	Virtual machine, 4 vCPU, 8 GB RAM, 300 GB storage, Ubuntu; NO GPU	Application services and production inference

Segmentation training ran at roughly 21 seconds per iteration on the processor. Production processing of a 6.24 GB raster took 6 minutes 31 seconds with peak memory use of 1,235 MB after resolution normalisation was applied; before it was applied, the process halted on memory exhaustion.

11.3 Parameters governing repeatability

Parameter	Value
Detection confidence threshold	0.30
IoU matching threshold	0.50
Fauna tile size	800 pixels, stride 700 pixels
Vegetation tile size	100 x 100 metres, 15 metre buffer
Vegetation reference resolution	4.8 cm/pixel
Segmentation architecture	Mask R-CNN, ResNet-101 backbone with FPN
Images per batch	2
Periodic evaluation	every 100 iterations
Bootstrap	B = 5,000, seed 20260818
Coordinate zone for area computation	UTM, determined automatically from data position

11.4 Data traceability

Every data file carries a SHA-256 fingerprint computed at ingestion and never altered afterwards. Every consolidated annotation retains a reference to its source file, and every automatically upgraded mask retains a marker of the method that produced it, so that silhouette annotations can be distinguished from those that remained rectangular.

11.5 Availability

Processing code, analysis scripts, and the fact-base file that is the source of every figure in this document are maintained in the project repository. Raw field data cannot be published openly because it contains precise coordinates of sensitive species; requests for access for scientific review may be addressed to the author.

12. Glossary

This list covers terms actually used in this document, not a general list. Indonesian equivalents are given because the companion Indonesian version of this report uses them, so that readers moving between the two versions can align the terminology.

12.1 Measurement and statistics

Term	Indonesian equivalent	Brief explanation
Ground truth	kebenaran dasar	Human-made annotation taken as correct, against which model output is compared.
True positive (TP)	positif benar	A prediction that matches ground truth.
False positive (FP)	positif palsu	A prediction unmatched to any ground truth. Requires care: a REAL object that happens not to be annotated also counts here.
False negative (FN)	negatif palsu	Ground truth the model failed to find.
Precision	Precision	The share of predictions that are correct. Falls when the model guesses wrongly - OR when the annotations are incomplete.
Recall	Recall	The share of ground truth successfully found. Unaffected by unannotated objects, so it remains valid even with incomplete annotation.
F1 score	F1	Harmonic mean of Precision and Recall.
Intersection over Union (IoU)	tumpang-tindih IoU	Intersection area divided by union area between prediction and ground truth. Ranges 0 to 1; 0.50 is the customary threshold.
Average Precision (AP, AP50, AP75)	AP, AP50, AP75	A summary of performance across confidence thresholds. The trailing number gives the IoU threshold: AP50 lenient, AP75 strict.
Greedy matching	pencocokan serakah	Predictions processed from the highest score; each immediately claims the best available match and is never revisited. The standard COCO evaluation approach.
Hungarian algorithm	pencocokan optimal global	The alternative to greedy matching, seeking the best overall combination of pairings. NOT used here, to keep figures comparable with the literature.
Confidence interval	selang kepercayaan	The range of plausible values for a measure given its sample size. A wide interval means little data.
Bootstrap	bootstrap	Estimating a confidence interval by repeatedly resampling the existing sample at random.
Paired difference	selisih berpasangan	Computing the interval on the DIFFERENCE between two models evaluated on the same sample. Stronger than comparing two separate intervals.
Statistically significant	signifikan secara statistik	The difference interval does not contain zero. It does NOT mean the difference is large or important.

12.2 Remote sensing and spatial data

Term	Indonesian equivalent	Brief explanation
Ground Sample Distance (GSD)	jarak sampel tanah	The real ground distance represented by one pixel, usually in centimetres.
Tile	petak	A small piece of a large image. Images are cut because models cannot process gigabyte-scale rasters in one pass.
Orthomosaic	ortomosaik	A composite image from many photographs, corrected so that scale is uniform.
Digital Surface Model (DSM)	model permukaan digital	The elevation of the topmost surface, including tree crowns and buildings.
Digital Terrain Model (DTM)	model medan digital	Ground surface elevation with vegetation and buildings removed.
Canopy Height Model (CHM)	model tinggi kanopi	DSM minus DTM, that is, true tree height. NOT available in this work because no DTM exists.
NDVI	indeks kehijauan NDVI	A measure of green vegetation density from satellite imagery.
Co-registration	ko-registrasi	Aligning two images from different sensors so that the same pixel points to the same place on the ground.
UTM zone	zona UTM	A metre-based coordinate system. Used because area and distance computation is not valid on degree-based coordinates.

12.3 Machine learning

Term	Indonesian equivalent	Brief explanation
Bounding box	kotak pembatas	A box enclosing an object.
Mask, instance mask	mask, siluet	The actual shape of an object rather than merely a box.
Instance segmentation	segmentasi instans	Drawing the shape of each object separately rather than simply marking a region.
Fine-tuning	penalaan lanjut	Continuing to train an already-trained model on new, more specific data.
Zero-shot	tanpa contoh	The model is asked to recognise something from a text description alone, with no example images.
Few-shot	dengan sedikit contoh	The model is given a handful of reference images before being asked to recognise.
Pseudo-labelling	pelabelan-semu	Using model output as additional training data after review.
Human-in-the-loop	manusia dalam lingkar	A workflow that requires human review before output is used.
Confidence threshold	ambang keyakinan	A cut-off on confidence value; predictions below it are discarded.
Non-maximum suppression (NMS)	penindasan non-maksimum	Discarding overlapping detections that refer to the same object.
Iteration	iterasi	One weight-update step during training.
Checkpoint	checkpoint	Model weights saved to disk at one point in training.
Early stopping	penghentian dini	Halting training when no improvement occurs after a number of evaluations.
Patience	kesabaran	How many evaluations without improvement are tolerated before early stopping.
Overfitting	pelatihan berlebih	The model memorises the training data so that performance on new data declines.
Backbone	tulang punggung	The part of the model that extracts features from the image, shared by all output heads.

12.4 Models and tools referred to

Name	Type	Role in this work
YOLO11n	Object detector	Production fauna detector until 20 September 2026. Now a comparator only; replaced for licence reasons (6.7).
D-FINE-N	DETR-family object detector, Apache-2.0 licensed	Production fauna detector since 20 September 2026, run through ONNX Runtime.
RF-DETR Nano	DETR-family object detector	Detector-comparison finalist; failed the speed gate.
ONNX Runtime	Model execution engine	Runs D-FINE-N without its training library.
Detectree2	Crown segmentation built on Mask R-CNN	Tree crown segmentation. In production use.
Mask R-CNN	Instance segmentation architecture	The architectural basis of Detectree2.
Detectron2	Modelling library	The framework that runs Detectree2.
Grounding DINO	Zero-shot detector	Baseline comparison and initial labelling aid, not the primary detection engine.
BioCLIP 2	Organism species classifier	Fauna species classification from image crops. Its underlying data is still very small.
GrabCut	Classical segmentation, not machine learning	Upgrading box annotations into crown silhouettes.
MegaDetector	Camera-trap wildlife detector	Studied only, not deployed; camera traps have not been procured.
Sentinel-2	Earth observation satellite	Image source for multi-temporal cover monitoring.
Chao2	Species richness estimator	Estimating inventory completeness from incidence data.

13. Appendix A - Foundations for the Non-Specialist Reader

13.1 How to use this appendix

Chapters 1 to 12 are written for readers already familiar with technical terminology, so much of the material is not explained from first principles. The three appendices that follow fill that gap without altering a single word of the preceding chapters - readers who already understand the material may skip them entirely.

The three appendices may be read as needed:

Part	Contents	Intended reader
Appendix A	Foundations from zero - what the problem is, what a model is, why humans are needed	Readers who have never dealt with artificial intelligence at all
Appendix B	A chapter-by-chapter walkthrough, explaining terms and figures where they appear	Readers working through the report who become stuck at a particular point
Appendix C	How to judge figures critically, including the traps that commonly mislead	Anyone who needs to ASSESS rather than simply believe claims based on artificial intelligence

Note: Appendices A and C can be read straight through. Appendix B is designed to be read alongside the chapter under consideration.

Cross-references use this document's own chapter numbers, for example "see Chapter 6.1".

CAUTION: These appendices explain WHAT was done and WHY. How the system was programmed is deliberately NOT explained and will never be included. The methodology is open so that anyone may scrutinise and re-test it - that is a condition of science. The implementation is closed, as is customary in applied engineering work.

13.2 The problem being solved

Baluran National Park covers tens of thousands of hectares. The officers who patrol it number in the dozens. To learn how many banteng remain, where they gather, or whether trees have been lost, the traditional method is to walk observation transects and record what is seen.

That method is scientifically sound but has three limitations that no amount of diligence can overcome: only a small fraction of the area is reachable, results depend on who is observing, and repeating it consistently every year is very expensive.

Unmanned aircraft - commonly called drones - can photograph thousands of hectares in a day. The problem then shifts: a single flight produces thousands of high-resolution photographs, and examining them one by one with human eyes takes longer than walking. This is where the computer is asked to help.

Note: Note the shift in the problem. The difficulty is no longer COLLECTING data but READING the data already collected. The entire methodology report is, at bottom, an answer to one question: how can that pile of imagery be read in a way that can be defended.

13.3 What "model" means in this report

The word "model" in the report does not mean a scale model or a sample. A model is a program that has been shown thousands of examples until it can recognise similar patterns in new images.

Imagine teaching someone to recognise banteng from aerial photographs. You would not explain "a banteng is dark brown, horned, so many metres long". You would show hundreds of photographs saying "this is a banteng", "this is not". Eventually that person can point them out unaided. They may not be able to put the reasons into words, but they can do it.

A model works in a roughly similar way. Two of its properties therefore matter from the outset:

- A model is only as good as the examples it has seen. A model that has only seen banteng on dry savanna will struggle with banteng in tall green grass.

- A model does not "understand" what a banteng is. It recognises patterns. Similar patterns - the shadow of a rock, a mound of earth, a buffalo - can be confused with it.

13.4 Annotation: the human work that cannot be replaced

Before a model can learn, someone must first mark the examples. Marking means opening an image and drawing a box around each banteng, or tracing a line along the edge of each tree crown. That work is called ANNOTATION, and its result is called GROUND TRUTH - the answer taken to be correct, against which everything is compared.

Annotation is tedious, slow, and expensive human work. It is also the source of several problems in this report - and that is precisely where one of its most important findings lies.

CAUTION: Remember this well: if the ground truth is wrong or incomplete, then EVERY assessment of the model is wrong too. It is not the model that is at fault but the measuring instrument. Chapter 7.2 of the report describes exactly this, and its consequences are large.

13.5 What "training" a model means

Training means showing annotated examples to the model repeatedly while it corrects itself a little each time it errs. One round of correction is called an ITERATION. The report mentions 600 iterations; that means the correction cycle was run 600 times.

Training from nothing would require millions of examples and expensive hardware. A cheaper route is therefore used: take a model already trained on millions of general images, then continue its training with Baluran imagery. It is like recruiting someone already trained to recognise animals in general, then teaching them Baluran's wildlife specifically. The term for this is FINE-TUNING.

Note: This is why the report states that adding and improving data yields far more than swapping between kinds of model. What decides the outcome is not how sophisticated the model is but how good the examples given to it are.

13.6 Two kinds of error, and why they differ in nature

A model can be wrong in two ways whose consequences differ sharply:

Kind of error	What it means	Practical consequence in the field
Missing	A banteng is in the photograph and the model does not find it	The population is undercounted; a threat goes unnoticed
False alarm	The model points at something that is not a banteng	Officers waste time checking something that is nothing

The two press against each other. Tuning a model to point more readily reduces missed cases but increases false alarms, and the reverse. No setting eliminates both at once.

Hence the two measures that always appear together in the report:

- RECALL answers: of all the banteng actually present in the annotation, what share was found? This falls when the model misses a lot.
- PRECISION answers: of everything the model pointed at, what share was correct? This falls when the model raises many false alarms.

The F1 figure is simply a way of summarising both into one number. It is useful for quick comparison but conceals which of the two errors is actually occurring.

13.7 Why measuring a model's ability is hard

Testing a model looks simple: give it images it has never seen and count how many it gets right. The difficulty lies in the word "right".

[1] Right according to whom? According to human annotation - which may itself be incomplete or mistaken.

2 How exact must exact be? If the model draws a box slightly offset from the annotated box, is that right? A rule is needed for how much overlap counts as a match. The report uses a half-area threshold, and states it openly because results change if the threshold changes.

3 Tested on which images? If tested on images very similar to its training images, the figures will look good but mean little for other conditions. The report states openly that this is precisely the limitation it has - see Chapter 7.1.

CAUTION: A model's performance figure NEVER stands alone. It always means "this much, MEASURED this way, ON data like this". A figure quoted without all three qualifications should not be trusted - no matter who is quoting it.

14. Appendix B - A Walkthrough of the Report, Chapter by Chapter

This part follows the report's order. Open the two side by side.

Document Status Statement

The report's opening page states that the system is still at proof-of-concept stage and must not be used as the sole basis for legal decisions. Statements of this kind are rare in technology marketing material, and placing it on the first page is a deliberate choice.

Note: How to read this: a report that states its own limits on the first page is generally more trustworthy than one that states them on the last page, or not at all.

Chapter 1 - Introduction

This chapter explains why the work was undertaken. Section 1.3 states that the contribution is "applied systems engineering, not algorithmic novelty".

What that means: every recognition engine used was built by others and is already well known. What is new is combining them into one system that actually runs on Indonesian data, complete with safeguards for data authenticity and explicit limits of authority. It is a modest but honest claim - and expert readers will respect it precisely because it does not overreach.

Chapter 2 - Literature Review

This chapter compares the work with research by others. Figures from Ball and colleagues (2023) are given as a benchmark for recognising tree crowns in tropical forest.

CAUTION: Note the caution box at the end of that chapter. It explains that those figures must NOT be placed directly alongside the figures in Chapter 6, because they are computed differently. This is a piece of honesty readers easily miss: the author could have placed them side by side and appeared favourable, but chose not to.

Chapter 3 - Study Area and Data Sources

This chapter details what data exists. Two points deserve a non-specialist's attention.

First, some data was REJECTED and the reason recorded, rather than quietly discarded. One data block was rejected because photographs from two camera types turned out not to match in count and position, so they could not be overlaid. Recording that rejection matters: it tells readers the data was not cherry-picked for a favourable result.

Second, the chapter states that the measure of "how many centimetres of ground each image point represents" is an estimate rather than an exact measurement. The reason is technical: the altitude recorded by the camera is measured from sea level, whereas what matters is height above the ground below - and Baluran's terrain is not flat.

Chapter 4 - Methodology by Pillar

The longest and most technical chapter. Three ideas are enough to grasp its core.

FIRST, two stages for fauna. The system does not ask "is this a banteng?" directly; it asks twice: "is there an animal here?", then "which animal?". The reason is practical: to add a new species later, only the second stage needs teaching, without redoing all the marking from the beginning.

SECOND, the size of image pieces is set by ground distance, not by image points. This sounds trivial but proved decisive. Large images must be cut into pieces before a model examines them. If pieces are defined by a count of image points, then on a sharper photograph one piece covers less ground, and a banteng appears far larger than the model is used to.

Note: The consequence is measurable and stated in the report: on one photograph the model found 45 objects; on another from a comparable location it found 297. Not because the model changed, but because the apparent size of the objects changed. Left unnoticed, the quality of the system would silently depend on whose photograph happened to be uploaded.

THIRD, cause-and-effect filtering before any accusation. When satellite imagery shows an area becoming less green, the system does NOT immediately call it logging. It first checks whether the location falls inside a licensed area, and whether the timing coincides with the dry season. This rule was born from a real event - see the note on Chapter 6.4 below.

Chapter 5 - Evaluation Protocol

This chapter sets out the rules of assessment BEFORE the results are presented. The order is deliberate: setting the rules after seeing the results opens the door to choosing whichever rules flatter most.

One term that may puzzle is "greedy matching". It works like this: when the model produces many guesses, the guess it is most confident about is examined first and immediately claims the annotation that best matches it. Once claimed, that annotation is unavailable to any other guess. It is called greedy because each step takes the best available at that moment, never revisiting the choice.

Note: This is not the optimal approach - other methods produce a better overall pairing. It is used anyway because it is the standard method used worldwide to assess systems of this kind, which keeps the figures comparable with other research. Choosing the better method would actually make the figures incomparable.

The chapter also states plainly that there is NO genuinely separate test set. The data used to train and the data used to test come from the same flight on the same day. This is a real weakness, and the report names it rather than waiting for someone else to find it.

Chapter 6 - Results

The core chapter. Its opening caution box states that several commonly reported measures are DELIBERATELY omitted. That is not an oversight, and the reason for each omission is given.

CHAPTER 6.1, wildlife detection. The figures to read: of 44 animals marked by humans in the test image, the second version of the model correctly found 39, with only two false alarms.

The figures in square brackets, for example [0.880-1.000], are called a confidence interval. They mean: "given this little data, the true value could plausibly lie anywhere in that range". The less data there is, the wider the range.

CAUTION: An interval that touches 1.000 does NOT mean the model is perfect. Quite the opposite: it is a sign that the data is too sparse to tell this model apart from a perfect one. The report says so openly, which matters - a great deal of technology marketing presents figures of this kind as proof of excellence.

The next table shows the difference between the older and newer versions, with the conclusion "Significant". That word has a specific meaning explained in Part 3.1 of this guide - and it does NOT mean "the difference is large".

CHAPTER 6.2, tree crown segmentation. Here only one measure is reported: recall. Precision and F1 are deliberately absent.

The reason is given in Chapter 7.2 and is the most important finding in the report: the tree crown annotation turned out to be incomplete. Many real trees were never marked by a human. As a result, when the model found those trees, the automatic assessment counted them as ERRORS - when the model was right and the humans had missed them.

Note: Recall is unaffected by this problem, because recall only asks "of those already marked, how many were found?" - trees never marked do not enter the count. That is why recall is still reported while the others are not.

CHAPTER 6.3, species identification. Note the caution that the Precision figure of 1.000 comes from TWO correct cases. A perfect score from two examples means nothing. The report includes it together with its warning, neither hiding the figure nor showing it off.

CHAPTER 6.4, satellite monitoring. This is an example of a NEGATIVE finding that is valuable precisely because it is negative. The greenness of the Baluran savanna fell sharply between April and August. Left to judge on its own, the system would have concluded that major logging had occurred. In reality it was an ordinary dry season.

Note: The value of this finding lies not in what succeeded but in the false accusation that was PREVENTED. A system that has not found something of this kind before deployment risks accusing people without basis.

CHAPTER 6.5, the confidence threshold. The model assigns a confidence value to each finding. Raising the acceptance threshold means fewer findings pass, but those that do are more convincing.

The table shows that trade-off quantitatively. What matters most: the report REFUSES to set the number, because the answer depends on how many candidates officers can actually review - and that is a decision for the managing institution, not a scientific one.

Chapter 7 - Limitations

This chapter lists ten known weaknesses. Reading the limitations chapter before the results chapter is a habit of experienced readers, and non-specialists are encouraged to copy it.

The three most consequential:

- There is no genuinely separate test data. All of it comes from one flight. This means the figures in Chapter 6 are valid for those photographic conditions, not a promise for other areas.
- Tree crown annotation is incomplete, which makes some measures impossible to interpret at all. It was found late, after all training had finished.
- The statistical testing method briefly differed between two parts of the work, and once made consistent, one conclusion CHANGED. The report records that change rather than erasing the trace of it.

Note: The third point deserves attention. The author could have quietly corrected it and presented only the final result. Recording that the conclusion once differed tells readers something important: results of this kind are sensitive to how they are tested.

Chapter 8 - Governance, Ethics, and Limits of Authority

This chapter affirms that the system produces HYPOTHESES, not decisions. Its output must not be the sole basis for law enforcement or for determining land boundaries.

Two practical points may interest a non-specialist. First, every data file is given a kind of digital fingerprint on entry; if even one image point is later altered, the fingerprint changes, so tampering becomes detectable. Second, the precise locations of rare wildlife and high-value trees are NOT shown to just any user, because disclosing locations can make poaching easier.

Chapters 9 to 12

Chapter 9 gives conclusions and planned next steps. Chapter 10 is the reference list. Chapter 11 is an appendix of technical details that allow others to repeat the testing. Chapter 12 is a glossary of terms.

Note: Chapter 11 deserves attention despite looking dull. It lists software versions, the random numbers used, and computer specifications. Its purpose is singular: to let others repeat and check the work. A report without details of this kind is difficult to verify.

15. Appendix C - Reading Figures Critically

This is the part most worth taking away from this document. It concerns how to assess claims based on artificial intelligence generally, not only the claims in this report.

15.1 "Significant" does not mean "large"

In everyday speech, significant means large or important. In statistics it means something entirely different: significant means the difference is consistent enough that attributing it to chance alone is not reasonable.

A small difference can be significant when there is plenty of data. A large difference can be NOT significant when there is little. The two words answer different questions: large answers "how far apart", significant answers "how sure are we the gap is real".

CAUTION: If marketing material describes a "significant" improvement without stating the amount of data behind it, the right question is: significant by which test, and tested on how many examples?

15.2 A figure without its sample size means nothing

This report contains a Precision figure of 1.000 - perfect. It comes from two cases. Two.

If someone tosses a coin twice and both come up heads, nobody concludes the coin always shows heads. Yet when the same thing is presented as a "100% success rate", many people accept it.

Note: A practical rule: whenever you see a figure for a system's ability, look for how many examples it rests on. If that is not stated, the omission is itself information.

15.3 A poor measurement may mean the measuring instrument is broken

This is the most important lesson in the report, and it reaches far beyond trees.

When the model's ability to recognise tree crowns was measured, the figures were poor. The natural conclusion is that the model is not very good. That conclusion turned out to be WRONG. On inspecting them one by one, most of the model's "errors" were real trees that humans had forgotten to mark.

The model was penalised for finding something that was there. What was broken was not the model but the answer key used to judge it.

CAUTION: Whenever a system is judged poor, two possibilities must be separated first: the system is genuinely poor, or the way it is being judged is wrong. Missing the second possibility discards systems that actually work - or, in the reverse direction, accepts systems that actually fail.

15.4 Two similar-looking figures may not be comparable

The report repeatedly refuses to compare figures that appear comparable at a glance. The reason is always the same: they were computed differently, or judged against different answers.

A concrete case within it: the older version of the system was judged against box-shaped marks, the newer version against marks tracing the crown edge. Guessing a box to match a box is plainly easier. The older version's higher figure therefore reflects an easier task, not greater ability.

Note: The test question that always applies: were these two figures measured the same way, on the same material, against the same answers? If any one of those differs, they must not be set side by side.

15.5 A report that names its failures deserves more trust

Common instinct says the opposite: a report full of successes feels more convincing. In science the reverse holds.

Every measurement carries uncertainty, every experiment has failures, and every system has limits. A report mentioning none of these does not thereby lack them - it merely does not mention them.

This methodology report records a rejected experiment, a process that failed on memory exhaustion, automatic marking that produced empty shapes, and one conclusion that changed after the testing method

was corrected. Their presence is not a sign of poor work but of genuine scrutiny.

15.6 Seven questions for testing any claim based on artificial intelligence

This list can be carried into any meeting where someone presents figures for an intelligent system - including when the person presenting is the author of this report.

1. Tested on how many examples? If the answer is in the dozens, every figure is provisional.
2. Is the test data genuinely separate from the training data? If both come from the same source, the figures tend to be too flattering.
3. Who produced the reference answers, and has their completeness been checked? Incomplete answers make a correct system look wrong.
4. Does the measure come with a range of uncertainty? A single figure without a range conceals how little data there is.
5. What FAILED? If there is no answer, the work has probably not been examined deeply enough.
6. What is this figure compared against, and was the comparison measured the same way?
7. What happens when the system is wrong, and who checks before a decision is acted upon?

Note: The seventh question matters most in real deployment. A system that errs but is always checked by a human is far safer than a system that errs less often but whose decisions are acted upon directly.

15.7 Closing

This guide is not intended to make its readers experts. Its aim is simpler and more useful: to let readers work through the methodology report to the end, understand what is claimed and what is not, and ask the right questions.

That ability to ask the right questions is what separates a reader who assesses from a reader who merely believes. For technology that will be used to manage publicly owned conservation areas, the difference matters.

Note: If after reading this guide you still find parts of the report hard to follow, the shortcoming lies in this guide, not in you. This guide will be improved.