University of Leeds — Laidlaw Scholars Summer Research Project

The Development of Automated Instrument-Use Detection for Surgical Tray Rationalisation: A Computer-Vision Pipeline

NIHR | National Institute for Health and Care Research

This project contributes to **SUSTAIN** (Sustainability and Simulation Theatre for Academia and Industry), led by the **NIHR HealthTech Research Centre in Accelerated Surgical Care** with **Leeds Teaching Hospitals NHS Trust**, the **University of Leeds** and the **Department of Health and Social Care**.

Rizlane Ouslimani

Supervisor

Dr Rory Turnbull

August 2026

Abstract

Surgical instrument trays frequently contain instruments that are rarely or never used, increasing sterilisation workload, cost and environmental impact. This study developed a YOLO11-based computer-vision pipeline to detect instrument presence and assess instrument use across seven dental surgical procedures, supporting evidence-based tray rationalisation.

The final merged dataset contained 20 instrument classes and was used to train and evaluate the object-detection model. The model achieved 76.3% precision, 88.9% recall, a mean average precision of 78.3% at an intersection-over-union threshold of 0.50, and 67.6% across thresholds from 0.50 to 0.95. Predictions were converted into normal frame counts, recording the number of detected instances, and binary presence sequences, recording whether each class was detected in each frame.

Instrument use was defined through a stable presence–absence–return transition. A minimum absence window filtered brief detection losses, while a surrounding-evidence layer required stable detections before and after each absence. Tool–case combinations with insufficient AI detection evidence were excluded rather than classified as unused. Across the seven cases, mean exact instrument-use count accuracy was 65.4%, while mean used/not-used accuracy was 71.9%. Exact-count accuracy ranged from 54.5% to 89.7%, and used/not-used accuracy ranged from 60.0% to 89.7%. Case 5 achieved the highest accuracy for both measures.

Exploratory tracking, which followed the same instrument across consecutive video frames, incorporated a polygonal tray region, appearance-based re-identification and a five-second removal-confirmation period. These measures reduced background false positives and unnecessary identity changes, although tracking accuracy could not be quantified without manually timestamped identities and use events.

The manual audit recorded no use for six instruments across the seven procedures: Coupland 3, Glasgow pattern rongeur, hammer, Mayo scissors straight, Spencer Wells artery forceps straight and Warwick James right. AI and manual records both indicated no use of the Glasgow pattern rongeur and hammer, providing the clearest initial evidence for clinician-led tray review. Additional cases, wider clinician participation and instrument-specific class labels are required before permanent tray changes are implemented.

Table of Contents

Part One: Research Output

Introduction

Climate change presents a growing threat to human health while healthcare delivery itself consumes substantial quantities of energy, materials and other resources. Global assessments suggest that healthcare is responsible for approximately 1–5% of total environmental impacts, depending on the indicator considered (Lenzen *et al.*, 2020). Within England, the National Health Service produced an estimated 25 million tonnes of carbon dioxide equivalent in 2019. Although this represented a 26% reduction from 1990, approximately 62% of the 2019 footprint remained associated with the healthcare supply chain, including pharmaceuticals, equipment and other purchased goods and services (Tennison *et al.*, 2021). Reducing unnecessary resource consumption is therefore integral to the NHS objective of reaching net-zero emissions by 2040 for directly controlled emissions and by 2045 for the wider emissions it can influence (Greener NHS, no date).

Operating theatres are especially resource intensive, using approximately three to six times more energy per unit area than hospitals overall. Surgical care also relies on equipment, packaging and sterilisation services, creating impacts through manufacture, transport, decontamination and waste management. Reducing unnecessary equipment processing may therefore lower both environmental and financial costs without changing the clinical procedure (MacNeill, Lillywhite and Brown, 2017).

Reusable instruments are assembled into standard trays so clinicians can respond to expected and uncommon needs. However, tray contents may grow over time through preference and historical practice. Once opened, unused instruments still require inspection, cleaning, repacking and sterilisation. A systematic review reported potential reductions of 19–89% in tray contents and consistent cost savings, but also found varied methods and considerable risk of bias. Objective procedure-specific data and clinician oversight are therefore needed (Eussen *et al.*, 2025).

Computer vision can automate this audit from images and video. Object detectors identify both an object's class and location. The You Only Look Once (YOLO) approach predicts classes and bounding boxes within one network, enabling rapid frame processing (LeCun *et al.*, 1998; Redmon *et al.*, 2016). This project used Ultralytics YOLO11n to balance detection performance with limited computing resources (Jocher and Qiu, 2024).

Previous studies show that deep learning can recognise surgical and dental instruments. Deol *et al.* (2024) reported high precision and recall for 11 surgical-instrument categories. Poomrittigul *et al.* (2026) reported mAP@0.50 of 95.9% and mAP@0.50–0.95 of 80.9% across 14 dental-instrument classes. Both studies noted the need for validation under realistic conditions with overlap, varied layouts and clinical backgrounds.

Procedural video is harder than standardised tray imaging. Dental instruments are small, reflective and visually similar, and their appearance changes with orientation, lighting, motion and occlusion. A fully obscured instrument produces an unavoidable false absence in that frame. Adjacent video frames are also highly similar, creating a risk of leakage if frames from one procedure enter both training and test sets.

This study was undertaken within SUSTAIN, a simulated operating-suite environment in

Leeds for evaluating sustainable clinical innovations. Tray rationalisation and asset tracking are core themes of the facility (NIHR HealthTech Research Centre in Accelerated Surgical Care, 2026).

The project aimed to develop and evaluate a YOLO11n pipeline for detecting and counting 20 dental-instrument categories in procedural footage. Detection, frame-based counting and exploratory tracking were assessed, and AI-inferred use was compared with manual records to examine feasibility for tray rationalisation. Detailed financial and life-cycle carbon calculations were outside the project scope.

Methodology

The pipeline comprised dataset preparation, YOLO11n training, prediction on 700 sampled frames, frame-based counting and temporal tracking. The final merged dataset contained procedural frames and supplementary instrument photographs. Predictions were evaluated on 100 frames from each of seven surgical cases. Counting measured class-level detections, while tracking attempted to maintain individual identities across frames.

3.1 Overall workflow

Table 1. Summary of the computer-vision workflow.

Stage	Input	Main process	Output
1. Dataset preparation	Procedural recordings and supplementary images	Frame selection, annotation and quality checks	One labelled YOLO dataset
2. Model training	Training and validation subsets	Transfer learning with YOLO11n	Final `best.pt` checkpoint
3. Prediction	700 sampled frames: 100 per case	Independent object detection and confidence export	Boxes, labels and confidence scores
4. Counting	Individual frames	Count retained detections by class	Per-frame normal-count and binary-presence tables
5. Tracking	Complete video sequence	BoT-SORT with custom identity and timing logic	Instrument events and time-dependent counts

3.2 Dataset preparation

The dataset combined frames from seven dental surgical cases with supplementary instrument photographs. Sampling limited adjacent near-identical frames, and frames were retained when the tray or working area was clear enough for reliable annotation.

Each visible instrument was assigned to one of 20 clinical categories and enclosed by a bounding box. Partly occluded instruments were labelled only when their class and boundaries remained identifiable; fully occluded instruments were excluded. Labels were exported in YOLO format as class identifiers and normalised box coordinates.

Figure 1. Example frame extracted from a surgical case. Background regions were masked in the dataset images to focus annotation and training on the instrument tray.

Supplementary photographs provided clearer views and additional orientations for instruments often obscured in procedural footage. They increased visual coverage but did not reproduce the same backgrounds, lighting and occlusion.

Figure 2. Supplementary photograph showing instruments that were frequently occluded or obstructed in the procedural footage.

3.2.1 Instrument taxonomy

The final taxonomy contained 20 dental-instrument categories reviewed with a practising dentist. Labels were kept consistent across annotation, training, prediction, counting and tracking. Table 2 gives representative images and clinical functions.

Table 2. Instrument images, clinical labels and functions.

Instrument image	Instrument label and clinical function
	**McIndoe scissors** Fine dissection and cutting of soft tissue.

Table 2. Instrument images, clinical labels and functions (continued).

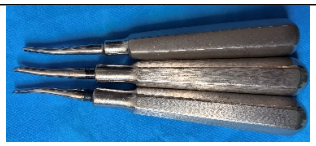
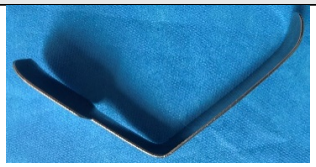
Instrument image	Instrument label and clinical function
	**Mayo scissors straight** Cutting sutures, dressings and dense tissue.
	**Coupland 1; Coupland 2; Coupland 3** Loosening teeth before extraction by severing the periodontal ligament.
	**Warwick James left; Warwick James straight; Warwick James right** Elevating and luxating tooth roots during extraction.
	**Cryers elevator left; Cryers elevator right** Removing fractured roots and retained root fragments.
	**Luxator** Cutting the periodontal ligament to allow atraumatic tooth extraction.
	**Minnesota retractor** Retracting cheeks, lips and soft tissues for improved access.
	**Lacs retractor** Retracting soft tissue to improve visibility of the surgical field.
	**Kilner retractor** Holding back soft tissue during oral surgery.
	**Glasgow pattern rongeur** Trimming and removing small amounts of bone.

Table 2. Instrument images, clinical labels and functions (continued).

Instrument image	Instrument label and clinical function
	**Kilner needle holder, gold-handled** Holding and guiding suture needles during wound closure.
	**Howarth's elevator** Elevating mucoperiosteal flaps from bone.
	**Buser** Harvesting, placing and contouring bone-graft material.
	**Mitchell trimmer** Trimming and contouring bone or soft tissue around the surgical site.
	**Hammer** Delivering controlled force to osteotomes or chisels.
	**Gillies dissecting forceps toothed** Grasping and stabilising tissue during dissection and suturing.
	**BP handle 3** Holding a surgical blade for making incisions.
	**Ball & socket towel clip** Securing surgical drapes in position.
	**Mosquito artery forceps curved** — clamping small blood vessels to control bleeding. **Mosquito artery forceps straight** — clamping small blood vessels and delicate tissue. **Spencer Wells artery forceps straight** — clamping blood vessels and controlling haemorrhage. **Fickling forceps** — grasping teeth, roots or small tissue during oral surgery.

Table 2. Instrument images, clinical labels and functions (continued).

Instrument image	Instrument label and clinical function
	**Dyson retractor** Retracting soft tissue to improve surgical access and visibility.

3.2.2 Data split and quality control

Labelled images were divided into approximately 70% training, 17% validation and 13% test subsets. Images and annotations were checked for missing pairs, invalid identifiers and inconsistent class names. The final checkpoint retained a legacy `tweezers` label with no distinct clinical mapping; it was excluded from the 20-category evaluation. Because frames from the same small case pool could appear across the development and later analyses, the case-based results are exploratory rather than an independent estimate of performance on unseen procedures.

3.3 YOLO11 model training

YOLO11n was selected to balance model size with the available CPU environment. Training used 640 × 640-pixel inputs for 50 epochs. Ultralytics produced loss curves, precision, recall, mean average precision, precision–recall curves and confusion matrices. The strongest validation checkpoint, **best.pt**, was used for prediction, counting and tracking.

Code 3.1. Core YOLO11 training configuration.

```
from ultralytics import YOLO

# Load pretrained YOLO11n weights.
model = YOLO("yolo11n.pt")

# Fine-tune the detector on the dental-instrument dataset.
results = model.train(
    data="<project_directory>/data.yaml",
    epochs=50,
    imgsz=640,
    device="cpu",
    project="<project_directory>/runs",
    name="dataset_training"
)
```

The YAML file contained the dataset paths and ordered class names. Project-relative placeholders replace personal file paths.

Predictions and confusion matrices were used to review repeated false positives, false negatives and confusion between similar instruments. Relevant annotations and training examples were checked before the final model was selected. Supplementary photographs improved class visibility but could not reproduce the occlusion, reflections, motion blur and background complexity of clinical footage.

3.4 Prediction on 700 sampled surgical frames

The final model was applied to 700 procedural frames: 100 randomly selected from each of seven cases. Equal sampling prevented longer recordings from dominating the analysis. Each frame was processed independently with **best.pt**; retained boxes, classes and confidence scores were saved for analysis using one common threshold.

Code 3.2. Prediction and confidence export for the 700-frame sample.

```
from pathlib import Path
from ultralytics import YOLO

MODEL_PATH = Path("<project_directory>/weights/best.pt")
EVALUATION_DIR = Path("<project_directory>/evaluation_frames")
OUTPUT_PROJECT = Path("<project_directory>/prediction_results")

# Apply the final checkpoint to the standardised evaluation sample.
model = YOLO(MODEL_PATH)
results = model.predict(
    source=EVALUATION_DIR,
    conf=0.30,
    imgsz=640,
    save=True,
    save_txt=True,
    save_conf=True,
    project=OUTPUT_PROJECT,
    name="final_700_frames"
)
```

The `save_txt` and `save_conf` options retained the class, box and confidence outputs required for later analysis. A confidence threshold of 0.30 was used for the 700-frame confidence analysis; the separate frame-counting pipeline used 0.50.

3.5 Frame-based instrument counting

Each frame was processed independently at a 0.50 confidence threshold. Two Excel tables were produced per case: normal counts recorded retained boxes per class, while binary presence recorded 1 when a class appeared and 0 otherwise. Rows followed chronological frame order and missing classes were recorded as zero. Normal counts supported audit, while binary sequences were used to assess instrument use.

Code 3.3. Frame-level normal and binary instrument-count outputs.

```
from collections import Counter
from pathlib import Path

import pandas as pd
from ultralytics import YOLO

CONFIDENCE_THRESHOLD = 0.50
model = YOLO(str(MODEL_PATH))
class_names = [model.names[i] for i in range(len(model.names))]
normal_rows, binary_rows = [], []

# Process frames chronologically so each row preserves video order.
for frame_index, frame_path in enumerate(frame_paths):
    result = model.predict(
```

```
        source=str(frame_path),
        conf=CONFIDENCE_THRESHOLD,
        save=False,
        verbose=False
    )[0]

    # Count retained boxes, then reduce the same counts to presence/absence.
    class_ids = result.boxes.cls.int().cpu().tolist()
    counts = Counter(class_names[class_id] for class_id in class_ids)
    metadata = {"Frame_Index": frame_index, "Frame_File": frame_path.name}
    normal_rows.append({
        **metadata,
        **{name: int(counts.get(name, 0)) for name in class_names}
    })
    binary_rows.append({
        **metadata,
        **{name: int(counts.get(name, 0) > 0) for name in class_names}
    })

# Export one normal-count and one binary-presence workbook per case.
pd.DataFrame(normal_rows).to_excel(
    OUTPUT_DIR / "normal_tool_counts_per_frame.xlsx", index=False
)
pd.DataFrame(binary_rows).to_excel(
    OUTPUT_DIR / "binary_tool_presence_per_frame.xlsx", index=False
)
```

3.5.1 Temporal assessment of instrument use

Detection alone does not indicate clinical use: an instrument can remain visible without being used or disappear briefly because of occlusion or detector instability. Use was therefore inferred from chronological binary presence.

An AI-inferred instrument-use event required a complete presence-absence-return pattern:

$$[\mathbf{1},\ \mathbf{1},\ \mathbf{1},\ \mathbf{1},\ \mathbf{1}] \ \rightarrow \ [\mathbf{0},\ \mathbf{0},\ \ldots,\ \mathbf{0}] \ \rightarrow \ [\mathbf{1},\ \mathbf{1},\ \mathbf{1},\ \mathbf{1},\ \mathbf{1}]$$

Absence had to last at least 10 frames, with detection in at least four of the five frames immediately before and after it. This rejected short occlusions and flickering boxes. Each qualifying absence block counted as one event. Initial absence and terminal disappearance were not counted because complete presence–absence–return evidence was unavailable. Combinations detected in fewer than 5% of frames, or lacking an output column, were labelled *Insufficient AI detection evidence* rather than unused.

Code 3.4. Temporal validation used to infer AI-detected instrument-use events.

```
import numpy as np

MIN_PRESENT_FRAMES = 5
MIN_ABSENT_FRAMES = 10
MIN_RETURN_FRAMES = 5
MIN_EVIDENCE_FRACTION = 0.75
MIN_DETECTION_FRACTION = 0.05

def run_lengths(sequence):
    """Return value, start, end and length for each consecutive block."""
```

```
    sequence = np.asarray(sequence, dtype=np.uint8)
    runs, start = [], 0
    for position in range(1, len(sequence) + 1):
        if position == len(sequence) or sequence[position] != sequence[start]:
            runs.append((
                int(sequence[start]), start, position - 1, position - start
            ))
            start = position
    return runs

def instrument_use_events(sequence):
    """Count validated presence -> absence -> return transitions."""
    sequence = np.asarray(sequence, dtype=np.uint8)
    events = []

    # Test each absence block against duration and surrounding evidence.
    for value, start, end, absent_length in run_lengths(sequence):
        if value != 0 or absent_length < MIN_ABSENT_FRAMES:
            continue
        if start < MIN_PRESENT_FRAMES:
            continue
        if end + MIN_RETURN_FRAMES >= len(sequence):
            continue

        before = sequence[start - MIN_PRESENT_FRAMES:start]
        after = sequence[end + 1:end + 1 + MIN_RETURN_FRAMES]
        if (before.mean() >= MIN_EVIDENCE_FRACTION and
                after.mean() >= MIN_EVIDENCE_FRACTION):
            events.append((start, end, absent_length))
    return events

# Low-coverage sequences are insufficient evidence, not non-use.
detection_fraction = float(sequence.mean())
if detection_fraction < MIN_DETECTION_FRACTION:
    ai_use_count = np.nan
    quality_flag = "Insufficient AI detection evidence"
else:
    ai_use_count = len(instrument_use_events(sequence))
    quality_flag = "Evaluated"
```

3.5.2 Comparison with manual instrument-use records

AI-inferred counts were mapped to clinical names and compared with manual records for the seven cases. Exact-count accuracy required identical totals; used/not-used accuracy required agreement only on whether use occurred at least once.

The signed count difference was calculated as:

$$AI - inferred\ instrument\ use\ count\ -\ manual\ instrument\ use\ count$$

Positive values indicate overcounting, negative values undercounting and zero exact agreement. Insufficient-evidence combinations were shown separately and excluded from conditional accuracy. When several instruments shared a YOLO class, AI activity was class-level only; separate manual names were retained diagnostically but did not represent independent AI estimates.

3.6 Temporal instrument tracking

BoT-SORT linked detections across frames and assigned track IDs, with `persist=True` retaining state. Analysis began at a configured 60-second point, and elapsed time was measured from the first processed frame. A global identity layer stored the class, last box and appearance features of missing tracks so later same-class detections could recover an earlier identity. Reassignment remained heuristic and could not prove physical identity.

3.6.1 Region-of-interest restriction

An interactive polygonal region of interest (ROI) limited analysis to the tray. The saved polygon could be reused with the same camera view, and a detection was retained when its box centre lay inside the boundary.

Code 3.5. ROI filtering using the bounding-box centre.

```python
import cv2
import numpy as np

def box_in_polygon(box, polygon):
    # Use the bounding-box centre as the ROI inclusion point.
    x1, y1, x2, y2 = box
    centre = ((x1 + x2) / 2, (y1 + y2) / 2)
    return cv2.pointPolygonTest(polygon, centre, False) >= 0

# Convert the saved tray boundary to the format expected by OpenCV.
roi_polygon = np.array(saved_polygon, dtype=np.float32)

# Ignore detections whose centres fall outside the tray region.
if not box_in_polygon((x1, y1, x2, y2), roi_polygon):
    continue
```

Figure 3. Polygonal region of interest used to restrict detection and tracking to the surgical tray.

3.6.2 BoT-SORT and appearance-based re-identification

BoT-SORT used appearance-based re-identification (ReID) and a 15-second track buffer. Its 0.35 confidence threshold was lower than the 0.50 counting threshold so weak detections caused by blur or partial occlusion could remain available for tracking.

Code 3.6. BoT-SORT configuration and frame-by-frame tracking call.

```
# Express the 15-second recovery period in frames for this video.
native_track_buffer = int(15 * video_fps)

# Configure motion, proximity and appearance matching.
tracker_config = {
    "tracker_type": "botsort",
    "track_high_thresh": 0.50,
    "track_low_thresh": 0.10,
    "new_track_thresh": 0.75,
    "match_thresh": 0.80,
    "proximity_thresh": 0.15,
    "appearance_thresh": 0.25,
    "gmc_method": "sparseOptFlow",
    "with_reid": True,
    "model": "auto",
    "track_buffer": native_track_buffer,
    "fuse_score": True
}

# persist=True retains tracker state between consecutive frames.
results = model.track(
    source=frame,
    persist=True,
    conf=0.35,
    tracker="auto_botsort.yaml"
)
```

3.6.3 Custom global identity reassociation

Global reassociation compared same-class candidates using spatial overlap, centroid distance and appearance. Appearance combined crop embeddings and HSV histograms with weights of 0.65 and 0.35. Scores measured similarity, not identity probability.

Reassociation rule	Acceptance criterion
Hard spatial match	IoU ≥ 0.08
Centroid-distance fallback	Distance ≤ 120 pixels
Soft spatial and appearance match	IoU ≥ 0.03 and appearance score ≥ 0.50
High-appearance lock	Appearance score ≥ 0.82, independent of position
Standard appearance fallback	Appearance score ≥ 0.55

Code 3.7. Simplified five-tier global-identity reassociation logic.

```
def find_best_match(missing_pool, box, class_id, embedding, histogram):
    # Compare only missing identities assigned to the same class.
    candidate, overlap, distance = best_spatial_candidate(
        missing_pool, box, class_id
    )

```

```
    # Prefer spatial continuity before consulting appearance features.
    if candidate is not None and overlap >= 0.08:
        return candidate
    if candidate is not None and distance <= 120:
        return candidate

    # Accept weaker overlap only when appearance also agrees.
    if candidate is not None and overlap >= 0.03:
        score = combined_appearance_score(
            embedding,
            missing_pool[candidate]["last_embedding"],
            histogram,
            missing_pool[candidate]["last_histogram"]
        )
        if score >= 0.50:
            return candidate

    # Use appearance alone only when spatial continuity is unavailable.
    candidate, score = best_appearance_candidate(
        missing_pool, class_id, embedding, histogram
    )
    if candidate is not None and score >= 0.82:
        return candidate
    if candidate is not None and score >= 0.55:
        return candidate
    return None
```

The listing is a shortened representation of the executable logic.

3.6.4 Temporal confirmation of removal

A missing detection was first marked pending. Reappearance within five seconds was treated as continuous presence; a longer absence generated a removal event. A short terminal gap remained unresolved, while stored identities allowed later returns to recover a previous global ID.

Code 3.8. Simplified five-second removal-confirmation logic.

```
# Convert the five-second occlusion allowance to a frame count.
grace_frames = int(5.0 * video_fps)

# Store a disappeared track as pending rather than removing it immediately.
for raw_id in disappeared_track_ids:
    global_id = active_id_map.pop(raw_id)
    missing_tools[global_id] = {
        "class_id": last_class[raw_id],
        "last_box": last_box[raw_id],
        "last_embedding": last_embedding[raw_id],
        "last_histogram": last_histogram[raw_id],
        "missing_since": frame_index,
        "removal_confirmed": False
    }

# Confirm removal only when absence exceeds the grace period.
for global_id, tool in missing_tools.items():
    absence = frame_index - tool["missing_since"]
    if not tool["removal_confirmed"] and absence > grace_frames:
        record_removal(tool["missing_since"], global_id)
        tool["removal_confirmed"] = True

# Do not force a short terminal gap into a removal at the video end.
```

```
if video_ended and absence <= grace_frames:
    record_event(frame_index, global_id, "unresolved_at_end")
```

3.6.5 Interactive inspection and visual output

Each class had a consistent colour. Global IDs remained visible, while selecting a box revealed its class and confidence to reduce label overlap. The interface also showed elapsed time, current counts and pending removals.

Code 3.9. Click-to-expand identity information.

```
# Keep detailed overlays hidden until the reviewer selects a tool.
selected_id = None

def handle_mouse_click(event, x, y, flags, parameter):
    global selected_id
    if event != cv2.EVENT_LBUTTONDOWN:
        return

    # Toggle the identity whose current bounding box contains the click.
    for x1, y1, x2, y2, tool_id in visible_boxes:
        if x1 <= x <= x2 and y1 <= y <= y2:
            selected_id = None if selected_id == tool_id else tool_id
            return
    selected_id = None
```

3.6.6 Tracking-data export

CSV and JSON exports recorded frame index, elapsed time, global ID, class and event type. A class-count history and final missing-instrument report were also produced.

3.7 Final analytical workflow

Counting provided direct per-frame class measurements. Tracking added identity and timing information but required more assumptions and parameters. Their suitability for tray rationalisation is compared below.

Results

4.1 Model training and validation performance

The final YOLO11n model was trained for 50 epochs using the merged dental-instrument dataset. Model performance was evaluated using the validation split, which contained 59 images and 1,608 annotated instrument instances.

Table 3 summarises the overall validation performance.

Table 3. Overall performance of the final YOLO11n model on the validation split.

Validation images	Annotated instances	Precision	Recall	mAP@0.50	mAP@ 0.50–0.95
59	1,608	76.3%	88.9%	78.3%	67.6%

Recall exceeded precision, indicating that most annotated instruments were detected but some false positives remained. The lower mAP@0.50–0.95 reflects the stricter box-overlap thresholds and shows that localisation was less consistent than class detection.

4.1.1 Training behaviour

Figure 4 shows loss and performance across 50 epochs. Training and validation losses generally fell, while precision, recall and mAP rose most strongly during the first 15–20 epochs before stabilising. Validation loss did not rise later, so the curves showed no clear divergence within training. Performance on procedural footage was assessed separately.

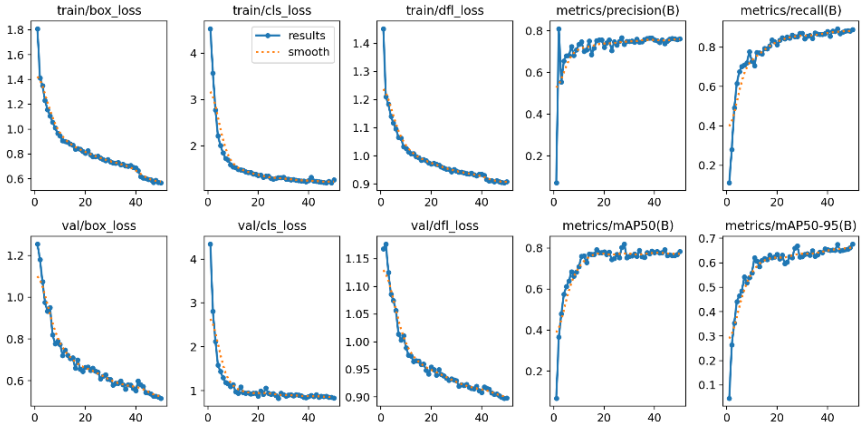


Figure 4. Training and validation losses, precision, recall and mean average precision across 50 epochs for the final YOLO11n model.

4.1.2 Class-level detection performance

Class-level performance was assessed using precision–recall curves and AP@0.50. Curves nearer the upper-right indicate better performance across confidence thresholds. Where clinically distinct instruments shared a YOLO label, metrics describe the grouped class rather than each instrument.

Table 4. Clinical instrument groups represented by each YOLO label and their class-level AP@0.50.

Clinical instrument class represented by the YOLO label	YOLO label	AP@0.50
Hammer	`a_mallet`	61.0%
Cryers elevators — left and right	`c_chs`	94.7%
Warwick James elevators — left, right and straight	`c_hook`	90.5%
Dyson retractor	`dyson_retra`	55.2%

Table 4. Clinical instrument groups and YOLO labels (continued).

Clinical instrument class represented by the YOLO label	**YOLO label**	**AP@0.50**
Howarth's elevator	`f_probe`	92.6%
Glasgow pattern rongeur	`forceps`	82.3%
Mayo scissors straight	`heavy_sci`	49.3%
Kilner retractor	`l_retra`	77.9%
Mosquito artery forceps curved and straight; Spencer Wells artery forceps straight; Fickling forceps	`lr_foreceps`	88.1%
Minnesota retractor	`minnesota_r`	65.2%
Mitchell trimmer	`mitchell_trim`	53.8%
Periosteal elevator Molt 9 and Buser	`n_probe`	82.9%
Kilner needle holder, gold-handled	`needle_h`	85.7%
Coupland elevators 1–3	`s_chs`	87.8%
Lacs retractor	`s_retra`	83.0%
BP handle 3	`sc_hand`	91.1%
McIndoe scissors	`scissors`	90.6%
Ball & socket towel clip	`sr_foreceps`	93.5%
McIndoe dissecting forceps and Gillies dissecting forceps toothed	`tweezer`	66.3%
Luxator	`vis`	74.5%

Overall mAP@0.50 was 78.3%. The strongest AP values were Cryers elevators (94.7%), ball and socket towel clip (93.5%), Howarth's elevator (92.6%), BP handle 3 (91.1%), McIndoe scissors (90.6%) and Warwick James elevators (90.5%). The weakest were Mayo scissors straight (49.3%), Mitchell trimmer (53.8%), Dyson retractor (55.2%) and hammer (61.0%). Likely causes include similarity, occlusion, reflections, orientation and unequal representation. Classes with few validation examples require cautious interpretation.

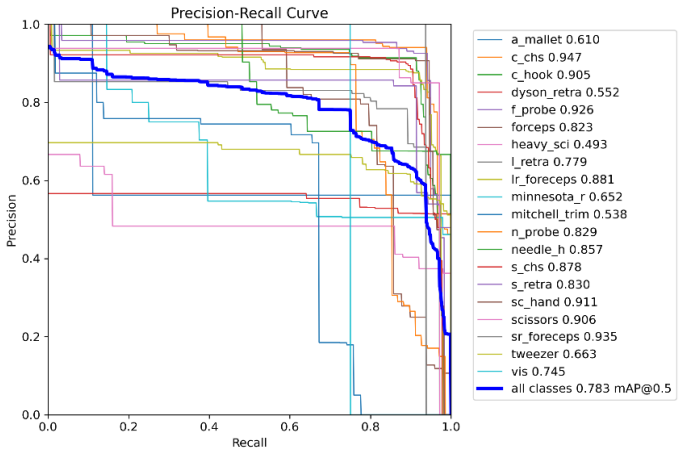


Figure 5. Class-level precision–recall curves for the final YOLO11n model. The thick blue line represents the mean across all classes (mAP@0.50 = 78.3%). Abbreviated labels correspond to the grouped clinical categories in Table 4.

4.1.3 Confusion-matrix analysis

In the normalised confusion matrix, diagonal cells show correct classifications, off-diagonal cells show class confusion, and the background row and column show missed instruments and false positives. Strong diagonal values indicate that detected instruments were usually assigned correctly; background errors were more common than direct class confusion.

Mitchell trimmer demonstrated the largest missed-detection proportion, with approximately 34% of its annotated instances assigned to the background. Higher missed-detection proportions were also observed for Luxator, the grouped Periosteal elevator Molt 9 and Buser category, and the Glasgow pattern rongeur. Background-related false-positive detections were particularly visible for Mayo scissors straight, the grouped dissecting-forceps category, Dyson retractor and Minnesota retractor.

Normalisation permits comparison between classes with different numbers of validation instances.

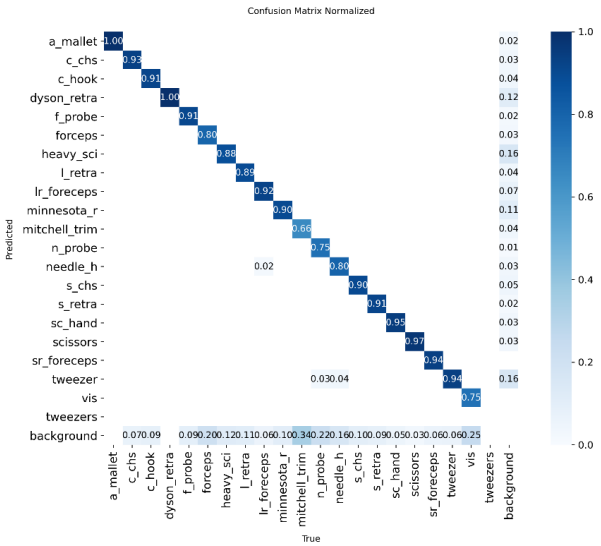


Figure 6. Normalised confusion matrix for the final YOLO11n model. Diagonal cells represent correct classifications; the background row and column represent missed and false-positive detections, respectively.

4.2 Prediction confidence scores

Figure 7 shows retained confidence scores. The central line is the median, each box contains the middle 50%, and the red line marks the 0.30 threshold.

Confidence varied between instrument categories, with most scores concentrated between approximately 0.40 and 0.60. Mitchell trimmer, McIndoe scissors, the Kilner needle holder and Minnesota retractor produced comparatively higher median scores. Lower medians were observed for the periosteal elevator Molt 9, Buser, the ball and socket towel clip and BP handle 3. The wide distributions for several classes indicate that prediction confidence varied considerably between frames.

Identical distributions for instruments sharing a YOLO label represent one underlying class. The plot excludes missed detections and does not measure accuracy directly. Moderate, vari-

able scores suggest uncertainty from occlusion, reflections, motion blur, orientation and overlap; manual comparison was therefore required.

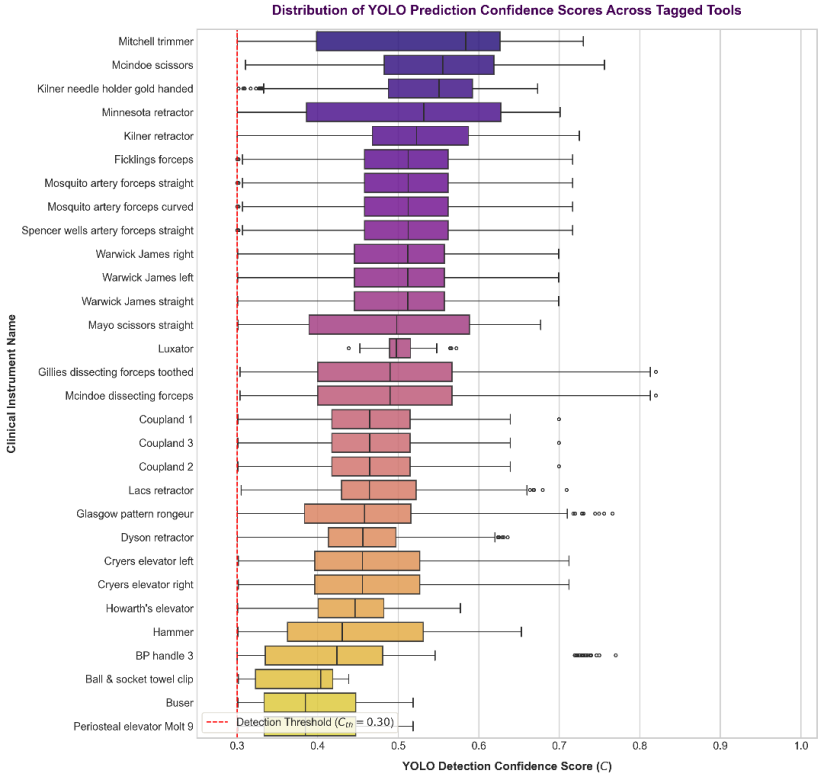


Figure 7. Confidence-score distributions for detections retained from 700 sampled surgical frames. The red dashed line marks the 0.30 threshold; instruments sharing a YOLO label have identical underlying distributions.

4.3 Instrument counting

YOLO was applied to the chronologically ordered frames from each surgical case. For every frame, the model produced a class label, bounding box and confidence score (Figure 8). Two Excel outputs were generated: a normal count recording the number of detections per class and a binary count recording presence (1) or absence (0). The binary sequences were subsequently used to infer instrument-use events.

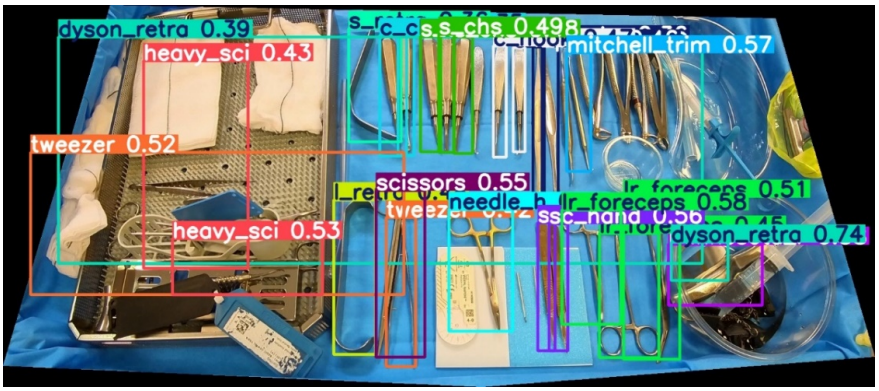


Figure 8. Example YOLO output showing detected instrument classes, bounding boxes and confidence scores within a surgical frame.

4.3.1 Accuracy by surgical case

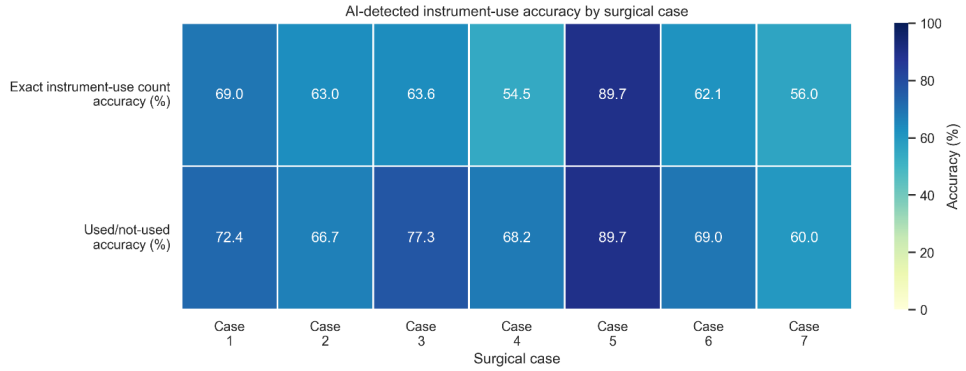


Figure 9. Exact usage-count and used/not-used accuracy by surgical case among combinations with sufficient AI evidence.

Two complementary measures were used to assess the AI's ability to infer instrument use from the frame sequence. **Exact usage-count accuracy** measures whether the AI inferred the *same number of use events* as the manual record for each instrument–case combination. It is therefore the stricter measure: the AI must identify both that the instrument was used and how many separate times it was used. **Used/not-used accuracy** measures whether the AI and manual record agreed simply on whether an instrument was used at least once. It assesses whether the system can distinguish instruments that were used from those that remained unused, without requiring the exact number of use events to match.

Used/not-used accuracy was higher than exact usage-count accuracy in six cases. This shows that the AI more often identified whether an instrument was used at all than it reproduced the exact number of separate use events. Case 5 performed best for both measures (89.7%). Cases 3 and 4 achieved 63.6% and 54.5% exact usage-count accuracy, respectively, while Case 7 had the lowest used/not-used accuracy (60.0%). Across cases, the unweighted mean was 65.4% for exact usage counts and 71.9% for used/not-used classification. Evaluable combinations for Cases 1–7 were 29/30, 27/30, 22/30, 22/30, 29/30, 29/30 and 25/30, respectively; these percentages are conditional on sufficient AI evidence and do not measure complete system coverage.

4.3.2 Overall AI and manual usage totals

Figure 10 shows total events rather than a mean. Agreement was observed for Fickling forceps, Coupland 2, mosquito artery forceps straight and several Warwick James instruments. Both methods recorded no use for the Glasgow pattern rongeur and hammer, making them the clearest candidates for clinician-led tray review.

The dominant pattern was undercounting. Examples include BP handle 3 (2 AI versus 11 manual), Kilner needle holder (1 versus 13), McIndoe scissors (3 versus 14) and curved mosquito artery forceps (2 versus 27).

Repeated values for instruments sharing one YOLO class are not independent instrument-level estimates.

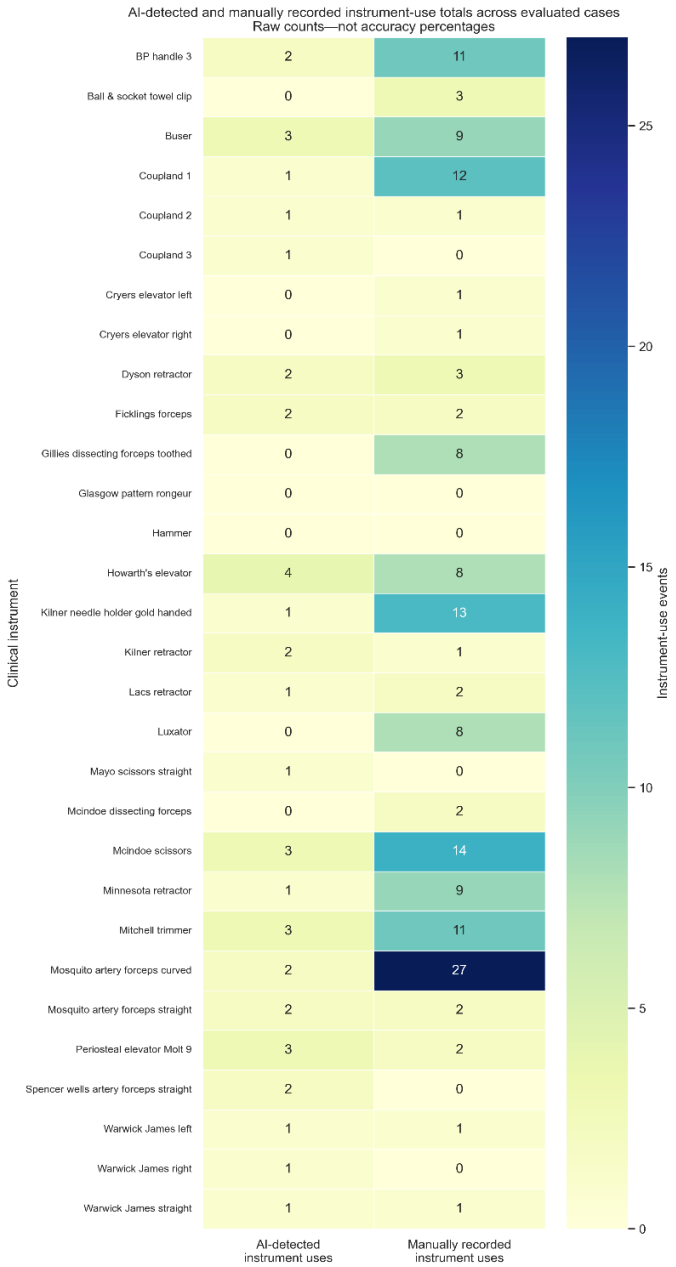


Figure 10. Raw AI-inferred and manually recorded instrument-use totals accumulated across evaluable surgical cases. Values are counts, not accuracy percentages.

4.3.3 Detailed error by instrument and case

The detailed heatmap reveals errors hidden by the overall percentages. The largest was -13 for curved mosquito artery forceps in Case 4. Other major undercounts were -8 for Buser and Luxator in Case 2, -7 for BP handle 3 in Case 2, and -6 for several instrument–case combinations. Positive errors were smaller, suggesting that validation suppressed false events but shifted the remaining error towards undercounting.

Grey cells marked Insuff. do not mean unused; they mark inadequate AI evidence and are excluded from conditional accuracy. Coverage must therefore accompany accuracy: exclusion avoids false non-use claims but can make apparent applicability seem stronger.

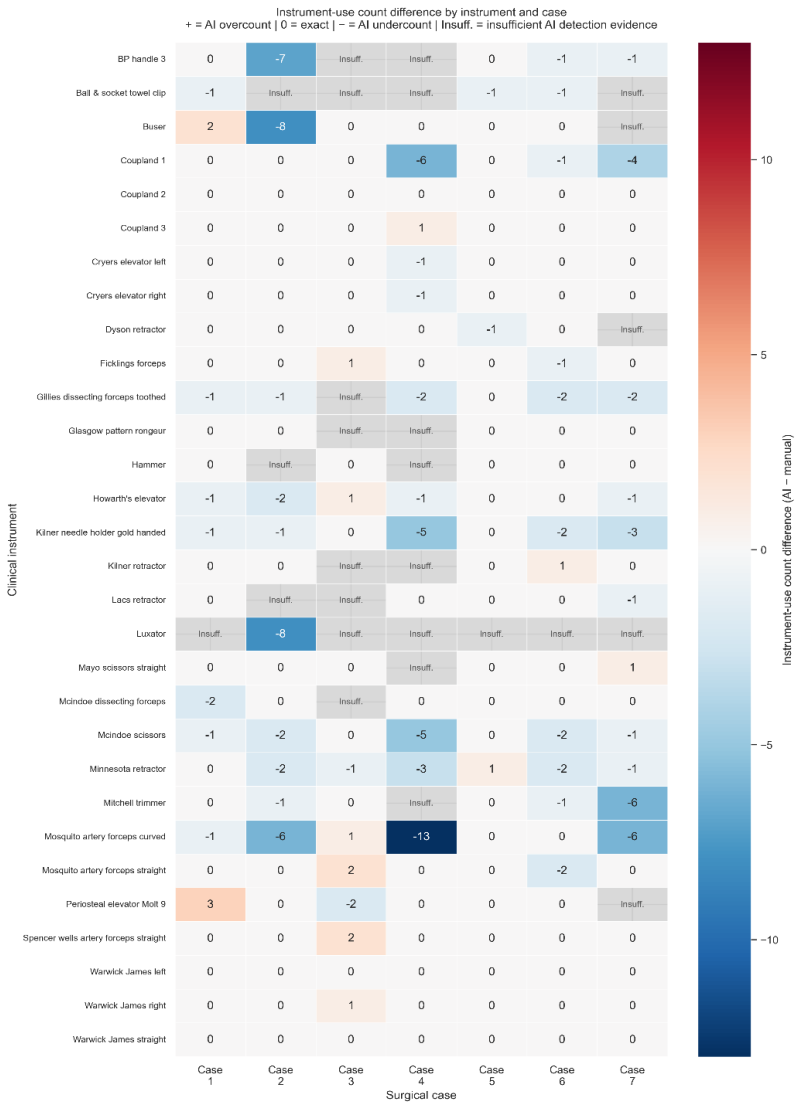


Figure 11. Difference between AI-inferred and manually recorded instrument-use counts by instrument and case. Positive values indicate overcounting, negative values undercounting, and 'Insuff.' indicates insufficient AI evidence.

4.4 Instrument tracking

The ROI-constrained tracker detected multiple instruments simultaneously and assigned each visible detection a class label and global tracking identity (Figure 12). Restriction to the tray removed many irrelevant detections originating from the surrounding clinical environment. Within the inspected footage, tracking remained visually stable during moderate tray movement and partial occlusion, while small bounding boxes could continue to represent partially visible instruments.

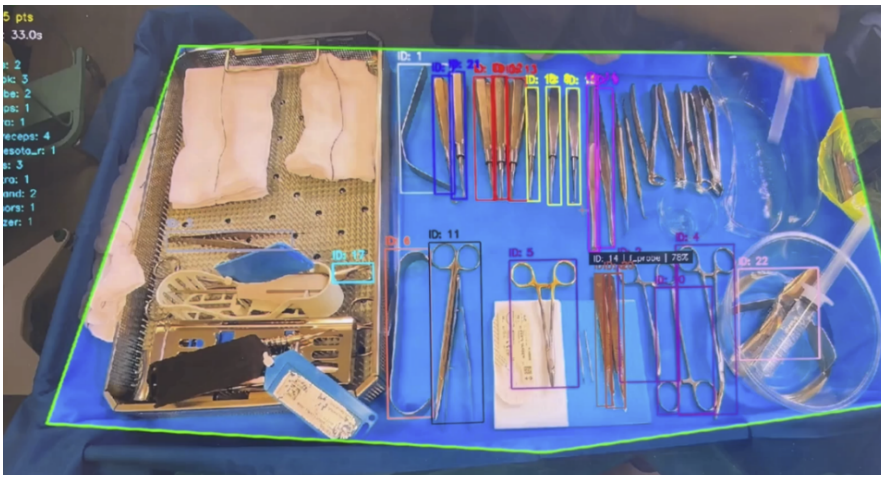


Figure 12. Example ROI-constrained tracking frame showing instrument classes, global IDs and live class counts.

Appearance-based ReID and the custom identity memory recovered earlier identities in a number of temporary-occlusion sequences. Figure 13 illustrates the visual audit used to examine whether IDs were maintained or recovered as instruments became obscured and visible again. This reduced, but did not eliminate, ID switching and inflation of the unique-instrument total.

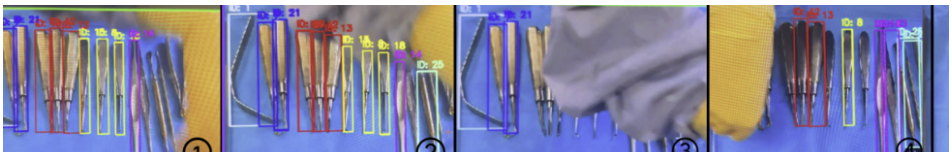


Figure 13. Sequential frames used to inspect tracking-ID continuity and reassociation following temporary occlusion.

The exported events were converted into a presence timeline (Figure 14). Coloured intervals indicate periods assigned to an identity by the tracker; gaps indicate pending or confirmed detection absence. Times are relative to the first analysed frame. The timeline added temporal context beyond the final usage count, and several intervals visually corresponded with removal and return in the video, but the event times were not independently annotated.

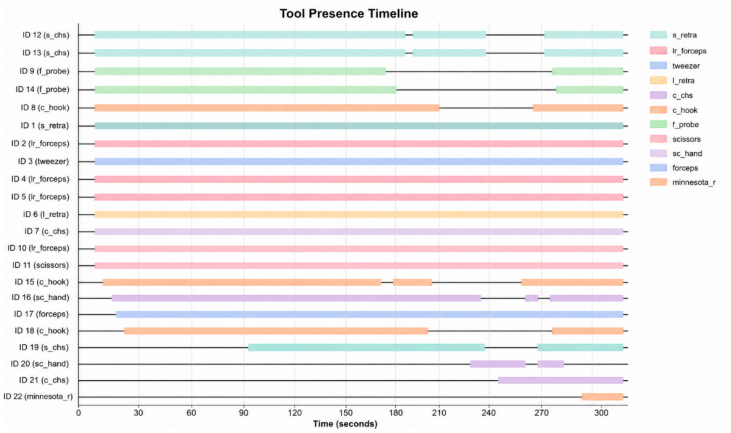


Figure 14. Example AI-detected instrument-presence timeline derived from tracking events.

Discussion

5.1 Instrument-based counting

The initial presence–absence–return method overcounted instrument use because temporary occlusion and unstable detections could imitate removal and return. The minimum absence duration removed short interruptions, while the surrounding-evidence layer rejected transitions not supported by stable detections. This explains the substantial improvement in Cases 3 and 4 without changing the underlying YOLO model.

The stricter method also produced genuine undercounting in some cases. A use event could be missed when an instrument was only intermittently detected before removal, remained occluded after return, was returned outside its original tray position, or disappeared at the end of the sequence. Reducing the thresholds could recover some events but would also reintroduce false positives; this is a sensitivity–specificity trade-off rather than a simple calculation error.

Shared class labels created an important structural limitation. Mosquito artery forceps curved, Mosquito artery forceps straight, Spencer Wells artery forceps straight, and Fickling forceps were recorded separately in the manual audit but represented by a shared YOLO forceps class. The AI could therefore detect use of the group, but could not identify which individual forceps had been used. Consequently, the apparent -13 undercount for Mosquito artery forceps curved in Case 4 should not be interpreted simply as 13 missed uses of that specific instrument; it largely reflects a mismatch between separate manual categories and the pooled AI class. Individual-instrument results for shared classes should therefore be interpreted cautiously. Separate annotation classes and model retraining would be required for reliable instrument-specific use counts.

Across all seven cases, Coupland 3, Glasgow pattern rongeur, hammer, Mayo scissors straight, Spencer Wells artery forceps straight, and Warwick James right were not recorded as used manually. These instruments are the strongest initial candidates for clinician review, particularly because the dentists had already considered removing some of them from the standard tray. In this context, the AI-supported audit provides evidence that can confirm a clinically motivated rationalisation decision rather than making the decision independently.

In contrast, rarely used instruments should not automatically be removed. Coupland 2, Fickling forceps, Kilner retractor, Lacs retractor, McIndoe dissecting forceps, Mosquito artery forceps straight, Periosteal elevator Molt 9, and Warwick James left and straight were used only once or twice in this dataset. However, rare use does not mean unnecessary use: for example, Fickling forceps may be required in emergency situations. Instrument selection may also vary between clinicians and procedures; the Kilner retractor was used infrequently partly because the operating dentist commonly retracted manually, whereas another clinician may use it more routinely.

These findings should therefore be treated as an initial, clinician-led tray rationalisation audit rather than a universal removal list. The dataset contained only seven procedures, mainly undertaken by two clinicians, so it may not represent wider clinical preferences or less common procedural needs. In addition, the training data and later frame-based analyses were derived from this same small case pool. Any split containing frames from the same procedure on both sides may benefit from case-specific visual similarity; future validation should therefore hold out complete procedures or capture sessions. As the project expands to approximately 50 cases with

a broader range of clinicians, the use patterns of rarely used instruments should be reassessed. Instruments consistently unused across this larger and more varied sample would provide stronger evidence for safe removal, while emergency or clinician-preference instruments may be better retained as accessible contingency items rather than included routinely on every tray.

5.2 Instrument tracking

The ROI addressed the mismatch between masked training images and the full clinical scene by suppressing many background false positives. Appearance-based ReID and global identity memory also improved continuity after short occlusions and reduced unnecessary new IDs. However, the ROI could not resolve confusion between similar instruments inside the tray, while severe occlusion, changed orientation or return to a different position could still prevent correct reassociation. The displayed unique-instrument total could therefore remain higher than the physical number of instruments.

The timeline should be interpreted as AI-detected presence and absence, not an independently validated record of clinical use. A temporarily lost bounding box could create a false absence even when the instrument remained on the tray, and instruments first detected later had no reliable earlier identity history. The estimated disappearance time was therefore not always the true time of removal for use. The five-second confirmation period also treated shorter disappearances as temporary occlusion, so genuine brief uses could be missed by the tracking-event count.

Transferability was further limited because the pixel-based thresholds depend on resolution, camera position and instrument size. Reassociation required class agreement, so temporary misclassification could block a correct match; conversely, appearance-only matching could merge similar same-class instruments. Future work requires manually timestamped events, additional cases and camera views, threshold calibration and a validated instrument-specific ReID model. Within the available project time, development was prioritised towards whether instruments were used and their total use frequency, which were more directly relevant to tray rationalisation. The timeline therefore remains an exploratory workflow-analysis output.

Conclusion

This project developed a YOLO11-based computer-vision pipeline to support dental surgical tray rationalisation by detecting instruments in procedural footage and inferring their use from changes in tray presence. Frame-level outputs were exported as normal detection counts and binary presence data, then converted into use events through a presence-absence-return rule. The additional temporal validation layer reduced false use events caused by brief occlusion and unstable detections, producing a more credible comparison with manual instrument-use records across seven surgical cases.

The final evaluation showed that the system more reliably identified whether an instrument had been used than the exact number of separate use events. Among combinations with sufficient AI evidence, the unweighted mean exact use-count accuracy was 65.4%, compared with 71.9% for used/not-used agreement. Performance varied between cases, reflecting differences in visibility, occlusion, detector stability and instrument-class definitions. The strongest limitations arose where multiple clinically distinct instruments shared one YOLO class: the system could detect

activity in the combined class, but could not assign that activity to a particular instrument. This explains some large apparent undercounts and demonstrates the importance of aligning model labels with the intended clinical question.

Exploratory BoT-SORT tracking, assisted by a tray polygon and appearance-based re-identification, demonstrated that individual detections could often be followed through partial occlusion and tray movement. The region of interest reduced detections outside the tray, while re-identification reduced some ID switching. However, missed boxes and imperfect identity recovery could still create false disappearance intervals, meaning that the tracking timeline is best interpreted as AI-detected presence and absence rather than a validated timestamped record of clinical use.

The most immediately useful contribution is therefore a scalable, clinician-supervised audit of instrument use. Instruments consistently unused across a larger and more diverse sample could be prioritised for review, while rarely used or emergency instruments should not be removed automatically. Extending the dataset to 50 or more cases, involving more clinicians and improving instrument-specific labels will determine whether the observed patterns generalise. In this form, computer vision can provide a practical evidence base for safer, data-informed tray rationalisation while retaining clinical judgement as the final decision-maker.

Part Two: My reflections

Impact of conducting research

At the start, most of the technical work was new to me. As a first-year chemical engineering student, I had never trained an object detector or developed counting and tracking systems, so the project initially felt intimidating.

The combination of healthcare, artificial intelligence and sustainability kept me motivated. Working with dentists showed me that a technically advanced model is useful only when its outputs answer a practical clinical question. This changed how I judged progress: a visually impressive output was not automatically the most useful one, and an apparently simple used/not-used result could have greater value for tray rationalisation than a more complex tracking display.

The main difficulties were learning unfamiliar theory and troubleshooting inconsistent results. Counting and tracking were more complex than detector training alone. I had to become comfortable with uncertainty, failed attempts and results that contradicted my expectations. I learned that research is iterative, that negative findings can reveal how a system behaves, and that limitations must be reported honestly rather than hidden by a headline accuracy value.

With support from my supervisor and clinical collaborators, I became more confident in investigating errors and making evidence-based decisions. I also became more willing to ask focused questions when I lacked expertise, while taking responsibility for turning the answers into practical changes. If I repeated the project, I would define clinical requirements and evaluation rules earlier and schedule more clinical checkpoints. This would reduce revisions and keep the work aligned with its intended use.

Leadership skills

Over the six weeks, I came to understand leadership less as directing other people and more as taking responsibility for direction, seeking the right expertise, responding constructively when evidence changes, and making decisions that others can understand and challenge. The project strengthened my communication, collaboration, critical thinking, problem-solving, project management and impact-evaluation skills.

Communication and collaborative mindset

Working with clinicians taught me to communicate across disciplines. At first, I had my own assumptions about which instruments should be prioritised for annotation and what an ideal technical output would look like. Discussions with the dentists gave me a clearer view of their practical objective: they primarily needed defensible evidence about whether instruments were used, rather than sophisticated outputs that did not directly inform tray composition. My priorities therefore shifted towards their clinical questions while remaining grounded in what could realistically be achieved within six weeks.

This required active listening rather than simply presenting my original plan. I asked how particular instruments were used, why rarely used instruments might still be essential, and what level of evidence would support a tray change. For example, learning that some instruments

are retained for emergencies or reflect operator preference prevented me from treating low use as an automatic recommendation for removal. I then translated those clinical distinctions into annotation, evaluation and reporting decisions.

I also learned to communicate technical limitations clearly. Instead of presenting a single accuracy figure without context, I distinguished exact use-count accuracy from used/not-used agreement and explained why shared YOLO classes could create apparent undercounts. This helped keep discussions focused on what the system could support safely. Leadership in this setting meant connecting different forms of expertise and making sure the technical work remained accountable to its clinical purpose.

Critical thinking and reflection

I learned not to accept model outputs without investigation. When the initial use estimates were inaccurate, I resisted the temptation to adjust parameters only until the percentages looked better. I examined the frame sequences and case-level error patterns, considered data quality, visual similarity, occlusion, class definitions and temporal limits, and asked why settings that helped one case sometimes harmed another.

This was particularly important when Cases 3 and 4 performed poorly under the first temporal method. A simple presence–absence–return transition treated short detection failures as instrument use. Comparing cases showed that the problem was not only the detector, but also the logic used to interpret its output. I therefore added a minimum absence duration and a surrounding-evidence layer requiring stable detections before and after an absence. The resulting improvement came from a better definition of a valid event, not from changing the underlying YOLO model.

The process also taught me when to stop iterating. Further relaxation of the rules could recover some undercounted events, but would reintroduce false positives. I learned to judge a solution through its trade-offs, intended use and evidence base rather than by pursuing a perfect headline number. That balance between ambition and methodological restraint became an important part of my decision-making.

Design thinking and problem-solving

I began the project with limited coding and computer-vision experience, so problem-solving often started with learning the theory behind an error. I learned new terminology, consulted documentation, tutorials, research papers and relevant open-source examples, and then tested whether those ideas addressed the behaviour visible in my own data. This was more effective than copying a solution without understanding its assumptions.

I separated detection, counting and tracking into related but distinct problems. Correct frame-level detection did not guarantee an accurate use estimate, and a stable class prediction did not guarantee a stable tracking identity. Testing each stage separately helped me locate failures and compare possible solutions. I used Excel exports and heatmaps to inspect temporal patterns, then revised the post-processing logic rather than retraining the model whenever the underlying detections were adequate.

The counting pipeline was the clearest example of thinking beyond my initial approach. My first presence–absence–return rule was too narrow and produced results that were not sufficiently

reliable for publication. I generated combinations of absence and stability windows, compared their errors across cases, and introduced a second validation layer that assessed the evidence surrounding each candidate event. This substantially reduced false events. The improvement showed me that good problem-solving requires revisiting the definition of the problem, not only tuning the first solution.

Tracking required a different form of experimentation. I added a polygonal region of interest to remove irrelevant background detections and explored appearance-based re-identification and a custom identity memory to recover IDs after occlusion. These changes reduced several visible errors but did not eliminate ID switching. Recognising that remaining limitation, and describing the timeline as exploratory rather than fully validated, was as important as producing the output itself.

Project management and prioritisation

The project included literature review, dataset preparation, annotation, training, prediction, counting, tracking and stakeholder communication. Managing these activities within six weeks required milestones, regular checks and conscious decisions about where further work would have the greatest value. I prioritised a complete and auditable counting pipeline because it answered the central tray-rationalisation question and could later be applied to many more cases.

My supervisors would have considered the validated counting work an appropriate endpoint, particularly because it supported the planned publication and conference abstract. However, I was ambitious about exploring what temporal tracking could add. Although they were understandably cautious about whether it could be completed within the available time, I chose to pursue it after securing the core counting output. I spent substantial time addressing background detections, occlusion, identity switching and visual presentation, and ultimately produced an informative tracking timeline and reassociation sequence.

This experience connected strongly with the Laidlaw value of ambition. For me, ambition did not mean ignoring constraints or claiming that an unfinished method was complete. It meant setting a demanding goal, remaining resilient when repeated attempts failed, and still protecting the reliability of the main study. I kept the tracking work explicitly exploratory and did not allow it to replace the better-validated counting results. This helped me distinguish productive persistence from perfectionism and showed me how to take a calculated research risk responsibly.

Impact measurement and research analysis

I learned that technical performance must be linked to practical impact. Bounding boxes alone were insufficient; outputs needed comparison with manual observations and interpretation for tray rationalisation. The used/not-used measure was particularly important because it matched the clinical decision more closely than an exact event count, while the coverage figures prevented excluded cases from making accuracy appear stronger than it was.

I also learned that evidence for impact must include context. Seven cases, mainly involving two clinicians, cannot represent every operator's preferences or every emergency requirement. Rather than presenting an automatic removal list, I framed the findings as evidence for clinician-led review and a starting point for a larger audit. This strengthened my curiosity, determination,

humility and confidence, and made me more aware that responsible leadership includes setting limits on what one's own results can justify.

Dissemination of my research

The findings have been shared with the SUSTAIN team and developed into this report, a research poster and conference submissions. With the support of my supervisor, research team and co-authors, my abstract was accepted for presentation at *From Ambition to Evidence: Advancing the Science of Sustainable Healthcare* in Malmö on 2 October 2026. Preparing the submission required me to reduce a complex six-week project to its central clinical question, most defensible results and principal limitations. It also showed me the value of teamwork in dissemination: the final abstract benefited from clinical, technical and academic perspectives that I could not have supplied alone.

I am also aiming to disseminate the work at a Laidlaw conference at Imperial College London, although this opportunity is not yet confirmed. Presenting there would allow me to exchange methods and ideas with Laidlaw scholars working in other fields and to consider how the same approach to evidence, iteration and sustainable practice might transfer beyond dental surgery. These opportunities have encouraged me to see dissemination not as the final administrative stage of research, but as a form of leadership through which findings are tested, improved and made useful to a wider community.

Future career, education and continued research plans

The project clarified my interest in biomedical engineering and research combining healthcare, sustainability and artificial intelligence. I now plan to strengthen my programming, machine-learning, computer-vision and data-analysis skills.

My chemical engineering background supports systematic thinking about processes, optimisation and sustainability. I hope to complement it through a master's degree in biomedical engineering or a related field and may later consider doctoral research.

The project showed that I can enter an unfamiliar field, learn independently and contribute without starting as an expert. It increased my confidence in interdisciplinary research.

References

Deol, E.S. *et al.* (2024) 'Artificial intelligence model for automated surgical instrument detection and counting: an experimental proof-of-concept study', *Patient Safety in Surgery*, 18(1), article 24. Available at: https://doi.org/10.1186/s13037-024-00406-y.

Eussen, M.M.M. *et al.* (2025) 'Reducing surgical instrument usage: systematic review of approaches for tray optimization and its advantages on environmental impact, costs and efficiency', *BJS Open*, 9(3), zraf030. Available at: https://doi.org/10.1093/bjsopen/zraf030.

Greener NHS (no date) *Delivering a net zero NHS*. Available at: https://www.england.nhs.uk/greenernhs/a-net-zero-nhs/ (Accessed: 31 July 2026).

Jocher, G. and Qiu, J. (2024) *Ultralytics YOLO11*. Version 11.0.0 [Computer software]. Available at: https://github.com/ultralytics/ultralytics.

LeCun, Y. *et al.* (1998) 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, 86(11), pp. 2278–2324. Available at: https://doi.org/10.1109/5.726791.

Lenzen, M. *et al.* (2020) 'The environmental footprint of health care: a global assessment', *The Lancet Planetary Health*, 4(7), pp. e271–e279. Available at: https://doi.org/10.1016/S2542-5196(20)30121-2.

MacNeill, A.J., Lillywhite, R. and Brown, C.J. (2017) 'The impact of surgery on global climate: a carbon footprinting study of operating theatres in three health systems', *The Lancet Planetary Health*, 1(9), pp. e381–e388. Available at: https://doi.org/10.1016/S2542-5196(17)30162-6.

NIHR HealthTech Research Centre in Accelerated Surgical Care (2026) *SUSTAIN: Sustainability and Simulation Theatre for Academia and Industry*. Available at: https://hrc-surgical.nihr.ac.uk/get-involved-2/sustain/ (Accessed: 10 August 2026).

Poomrittigul, S. *et al.* (2026) 'Deep learning-based object detection of restorative dental instruments with potential implications for workflow automation and infection control in dental supply units', *Scientific Reports*, 16, article 1168. Available at: https://doi.org/10.1038/s41598-025-30774-z.

Redmon, J. *et al.* (2016) 'You only look once: unified, real-time object detection', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. Available at: https://doi.org/10.1109/CVPR.2016.91.

Tennison, I. *et al.* (2021) 'Health care's response to climate change: a carbon footprint assessment of the NHS in England', *The Lancet Planetary Health*, 5(2), pp. e84–e92. Available at: https://doi.org/10.1016/S2542-5196(20)30271-0.