Rizlane Ouslimani | Supervisor: Dr. Rory Turnbull

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

LAIDLAW SCHOLARS

UNIVERSITY OF LEEDS

Background

Reusable dental trays may contain rarely used instruments, yet all instruments on an opened tray require cleaning and sterilisation, consuming resources. This SUSTAIN project assessed whether computer vision could support clinician-led tray rationalisation by identifying instrument presence and inferred use.

Aims

01 Develop and evaluate a YOLO11n pipeline for detecting and counting 20 instrument categories across seven surgical cases.

02 Assess its suitability as a scalable tool for instrument-use audits.

Methodology

Note: Instrument tracking was evaluated separately as a complimentary approach for more detail.

01 Training YOLO11n on **561** annotated images for **50** epochs → prediciton on new frames.

02 Recording class counts and binary instrument presence per frame.

03 Converting frame detections into use events using confirmed presence-absence-return transitions..

04 Calculating exact instrument-use count accuracy and used/not-used accuracy against the manual audit.

Key Findings

Validation performance:

- Dataset: **561** annotated images, **20** instrument classes, divided into a **70%/ 17%/ 13%** training-validation-test split.
- Precision: **76.3%** | Recall: **88.9%** | mAP@0.50: **78.3%** | mAP@0.50-0.95: **67.6%**

Prediction:

- Predictions above **0.30** confidence were retained.
- Most classes achieved median confidence of **0.50-0.70**.
- **Similar** or **occluded** instruments showed lower confidence.
- **Cases 3 and 4** provided case-level evaluation on procedures not represented during training.

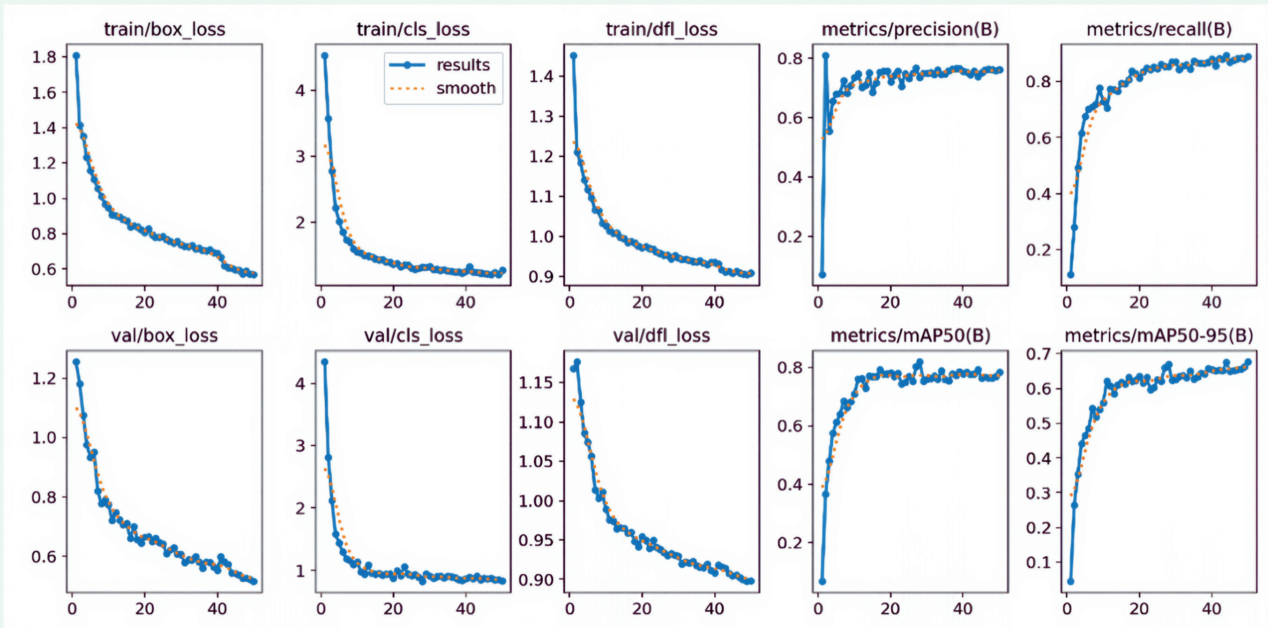

Figure 1. *Training and validation losses, precision, recall and mean average precision across 50 epochs.*

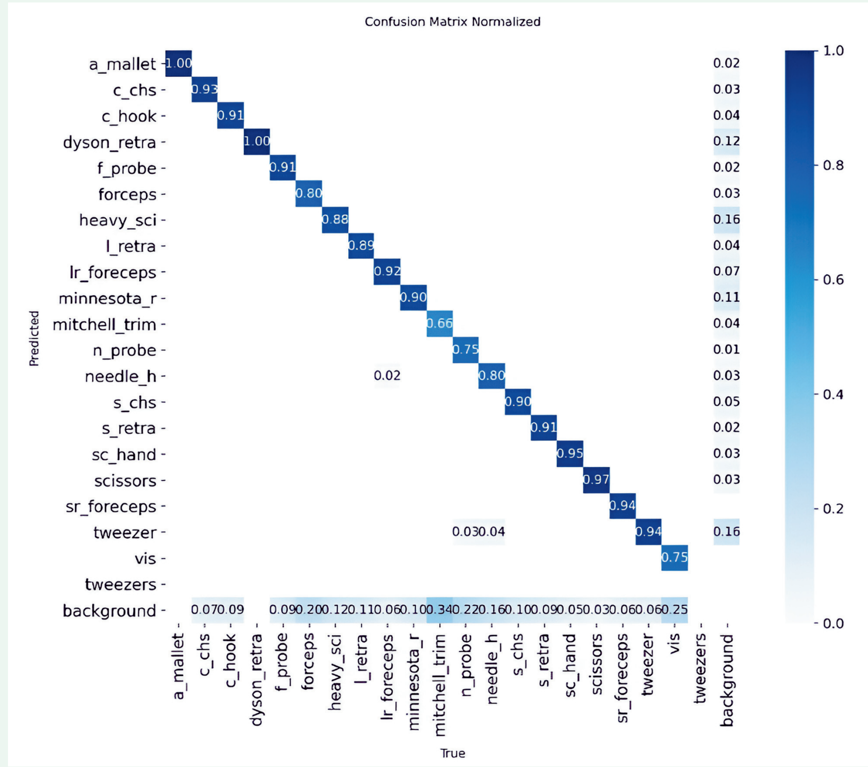

Figure 2. *Normalised Validation Confusion Matrix.*

Counting and accuracy evaluation:

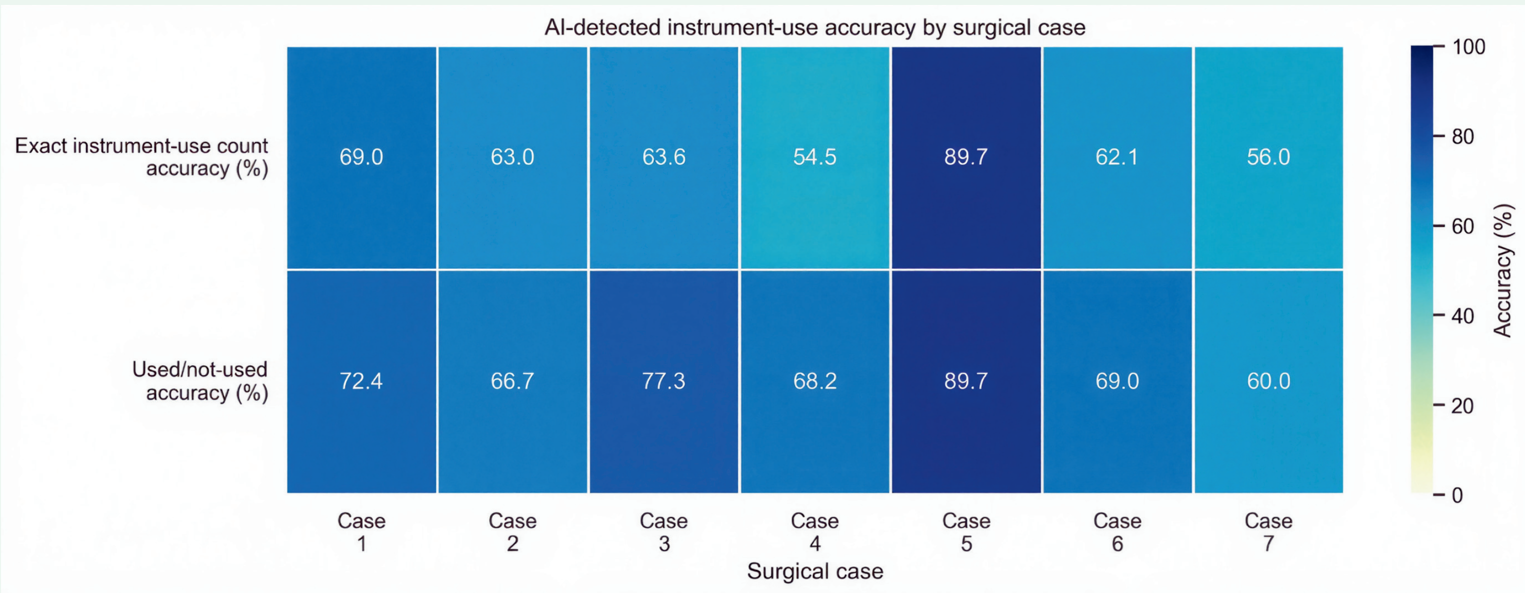

Figure 3. *Exact-count and at-least-once use agreement by case. Only instruments detected in more than **5%** of that case's frames were included.*

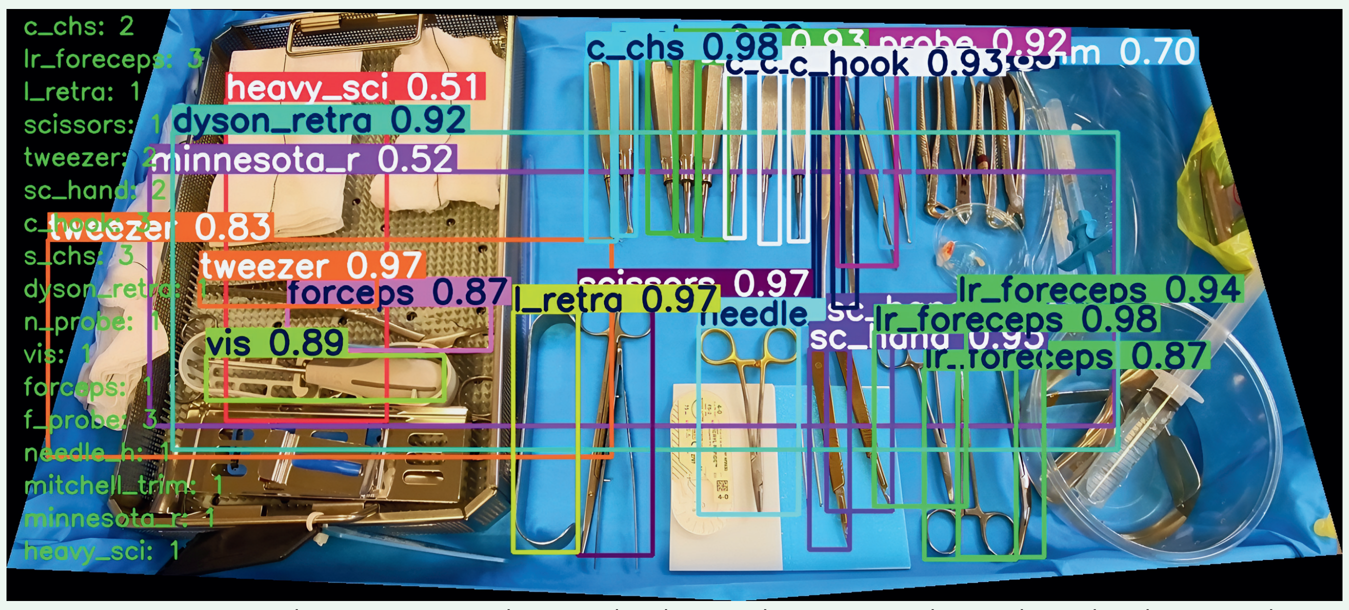

Figure 4. *Example YOLO output showing the detected instrument classes, bounding boxes and confidence scores within a surgical frame.*

- Used/not-used accuracy was **71.9%** versus **65.4%** for exact counts, showing better performance for screening potentially unused instruments than reconstructing every use event.
- Manual audit found **6 unused** and **9 rarely used** instruments. Both AI and manual review found no use of the Glasgow rongeur and hammer, supporting **clinician review**, not automatic removal. Rarely used instruments may also reflect **emergency needs** or **individual preferences.**
- Accuracy included only instruments detected in **>5%** of frames. **Occlusion, missed detections** and especially **grouped AI classes** caused **undercounting** and limited reliability, so outputs should support clinician-led decisions.

Instrument tracking:

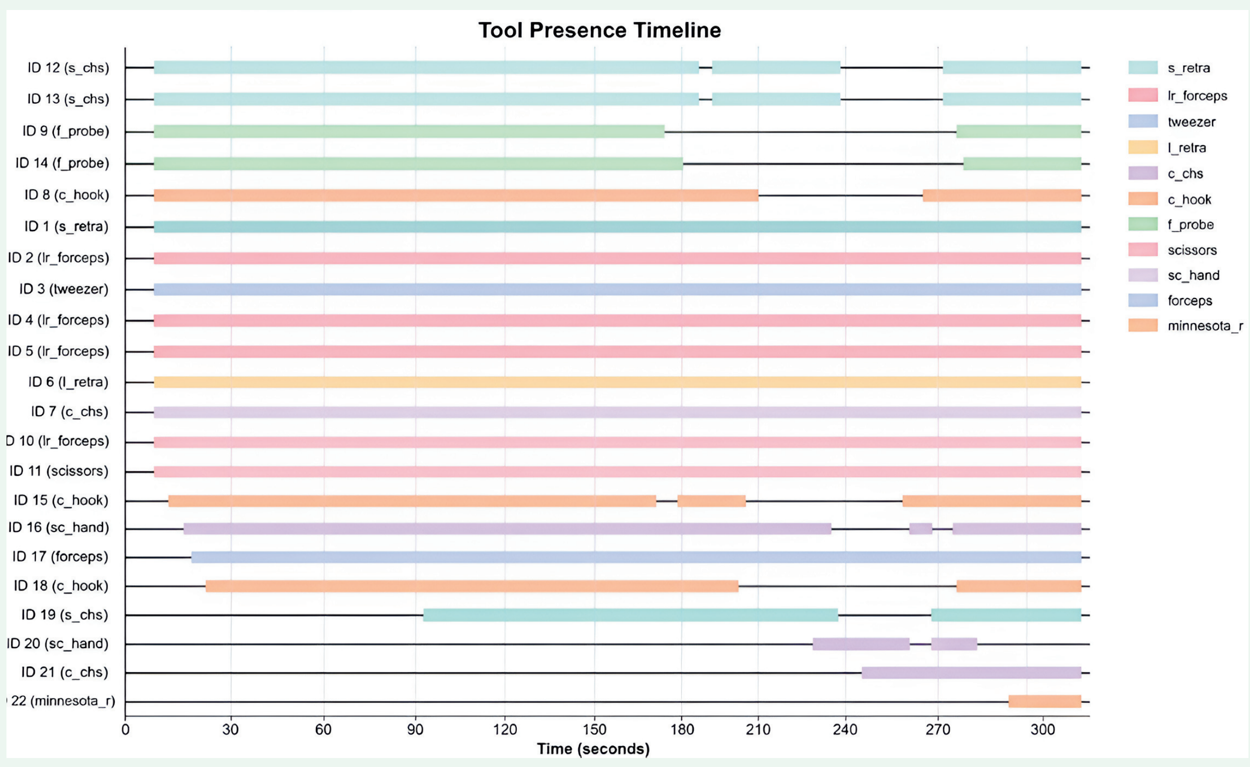

Figure 5. *AI-detected instrument-presence timeline derived from tracking events.*

Figure 6. *Tracking outputs: **(a)** ROI-constrained detections with instrument IDs; **(b-f)** sequential frames showing ID continuity across temporary occlusion.*

- The **green boundary** defines the tray region of interest, while **coloured boxes and IDs** track detected instruments. Sequential frames illustrate whether identities were **maintained** through temporary occlusion.
- SCAN to **view annotated footage** from **Case 2** one of seven analysed procedures (plus references and researcher profiles).

- A tray region of interest (ROI) reduced detections from the surrounding clinical environment, while re-identification (ReID) helped **recognise the same instrument after brief occlusions.**
- **Severe occlusion, changes in orientation and similar-looking instruments** could still produce false absences or duplicate identities.
- The timeline therefore shows **AI-detected instrument presence**, not validated clinical use or exact removal times, and remains an exploratory output requiring further validation.

Conclusion

- Used/not-used accuracy was **higher** than exact-count accuracy and may be **sufficient** at this stage to support tray-rationalisation screening, where the key question is whether an instrument was used at all. However, **undercounting** must be reduced.
- **At scale**, it could reduce manual auditing and help clinicians identify consistently unnecessary tray contents, **reducing avoidable cleaning, packaging** and **sterilisation**—and therefore saving water, energy, materials and staff time.

Next Steps

- Assign **separate annotation classes** to clinically distinct instruments where sufficient training images are available.
- Evaluate a **larger dataset** of **40-50 procedures** involving **more clinicians**, as the seven current cases mainly reflect the preferences of two clinicians.
- Further **validate** tracking for temporal workflow analysis.

Selected references: Eussen et al. (2025); Deol et al. (2024); Poomrittigul et al. (2026); MacNeill et al. (2017); Lenzen et al. (2020); Jocher & Qiu (2024); full references via QR.