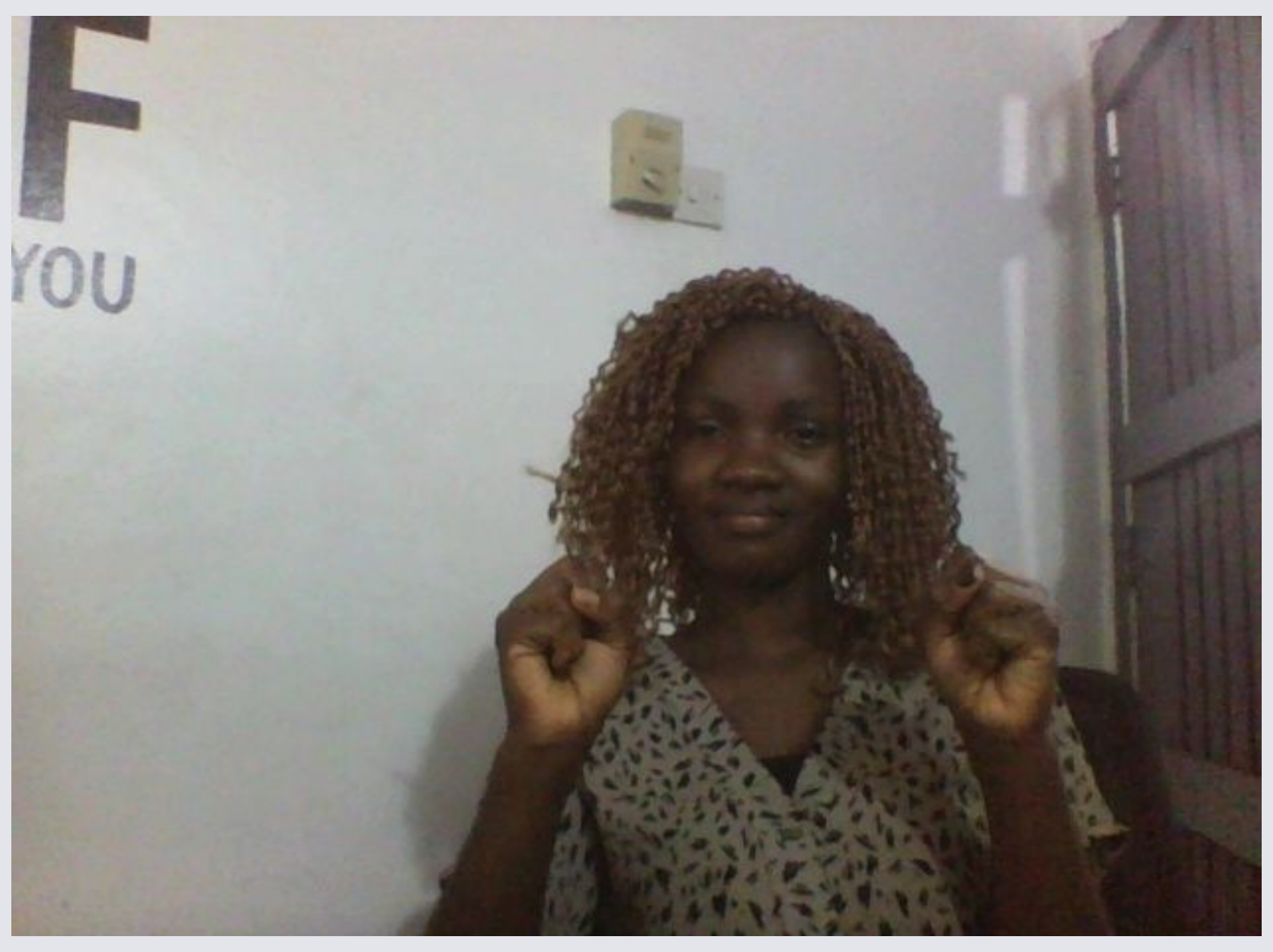
Towards an Inclusive Society: Sign-to-Speech Modeling for Sign Language Understanding

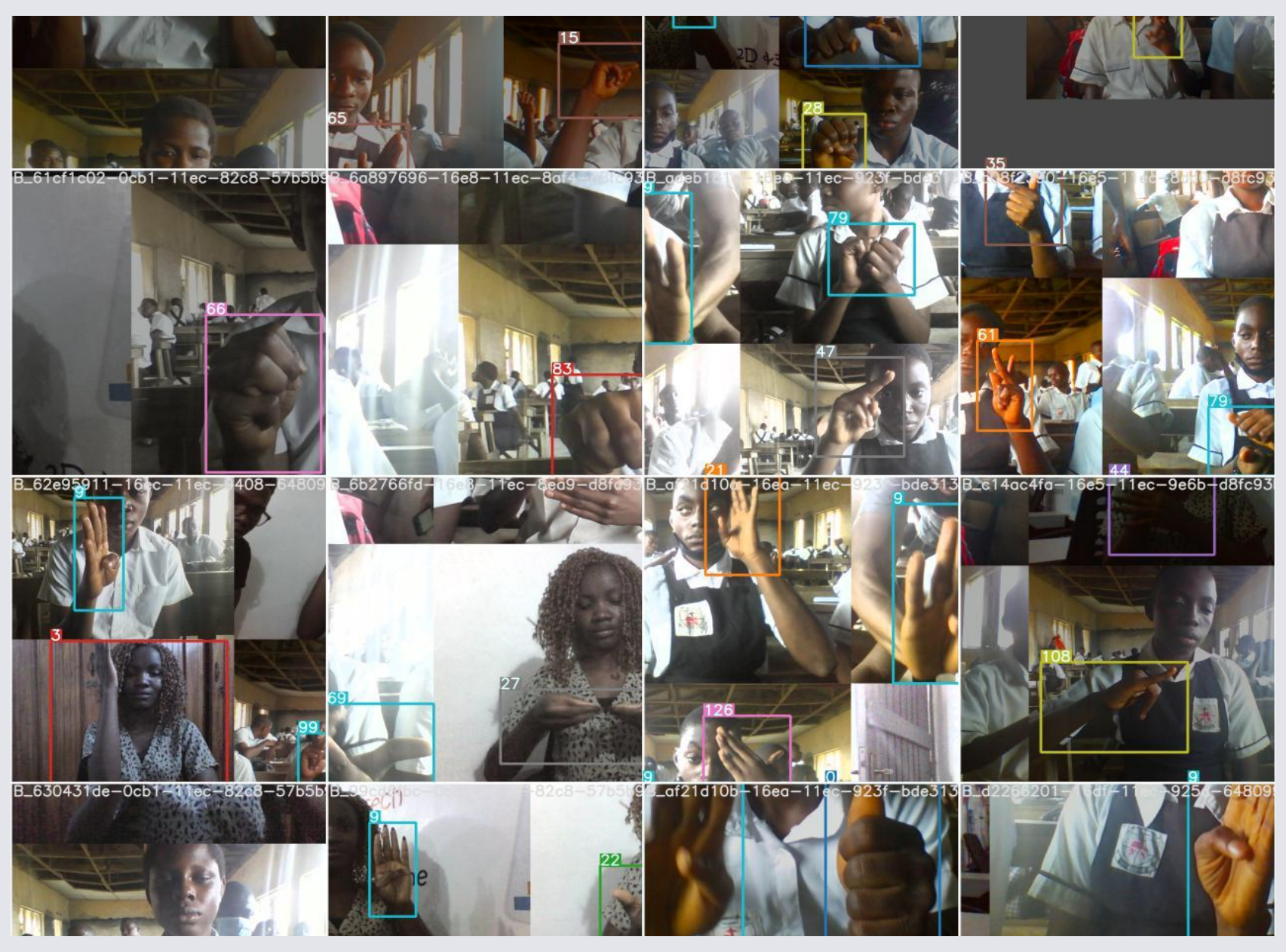
Steven Kolawole[1][2] Opeyemi Osakuade[3] Nayan Saxena[1][4] Babatunde Kazeem Olorisade[5]

ML Collective[1] Federal University of Agriculture Abeokuta[2] Data Science Nigeria[3] University of Toronto[4] Cardiff Metropolitan University[5]



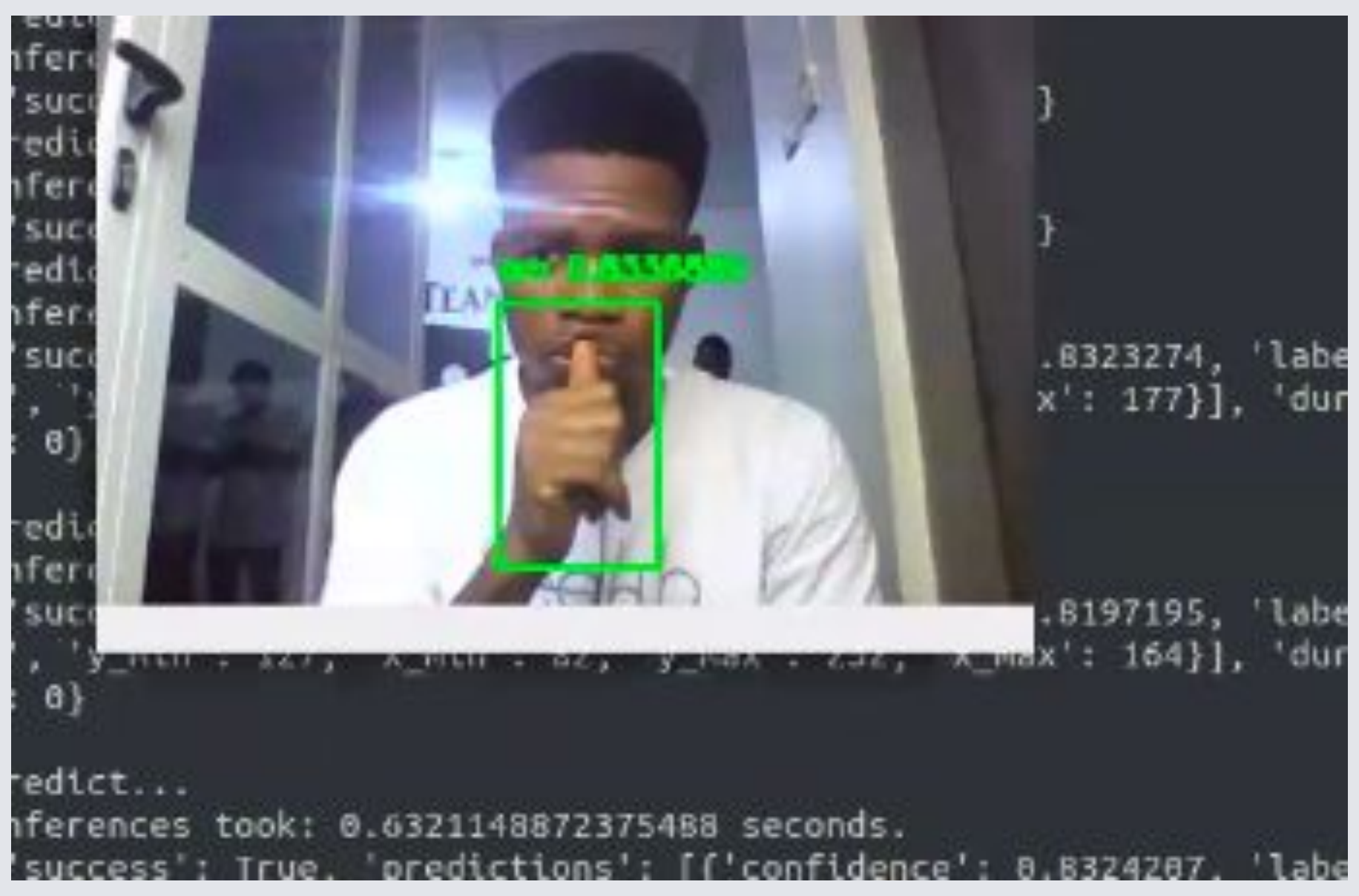
The 1st batch of the dataset was created by Amanda Bibire of the Ogun State Broadcasting Corporation.

After seeing a proof of concept prototype, teachers and students from 2 special education schools in Lagos and Abeokuta were very enthusiastic to help create the 2nd and larger batch of the dataset.



What we have afterwards is a widely-dispersed dataset of 20+ individuals captured in **diverse backgrounds and lighting conditions***.

A sample batch of test data true labels.

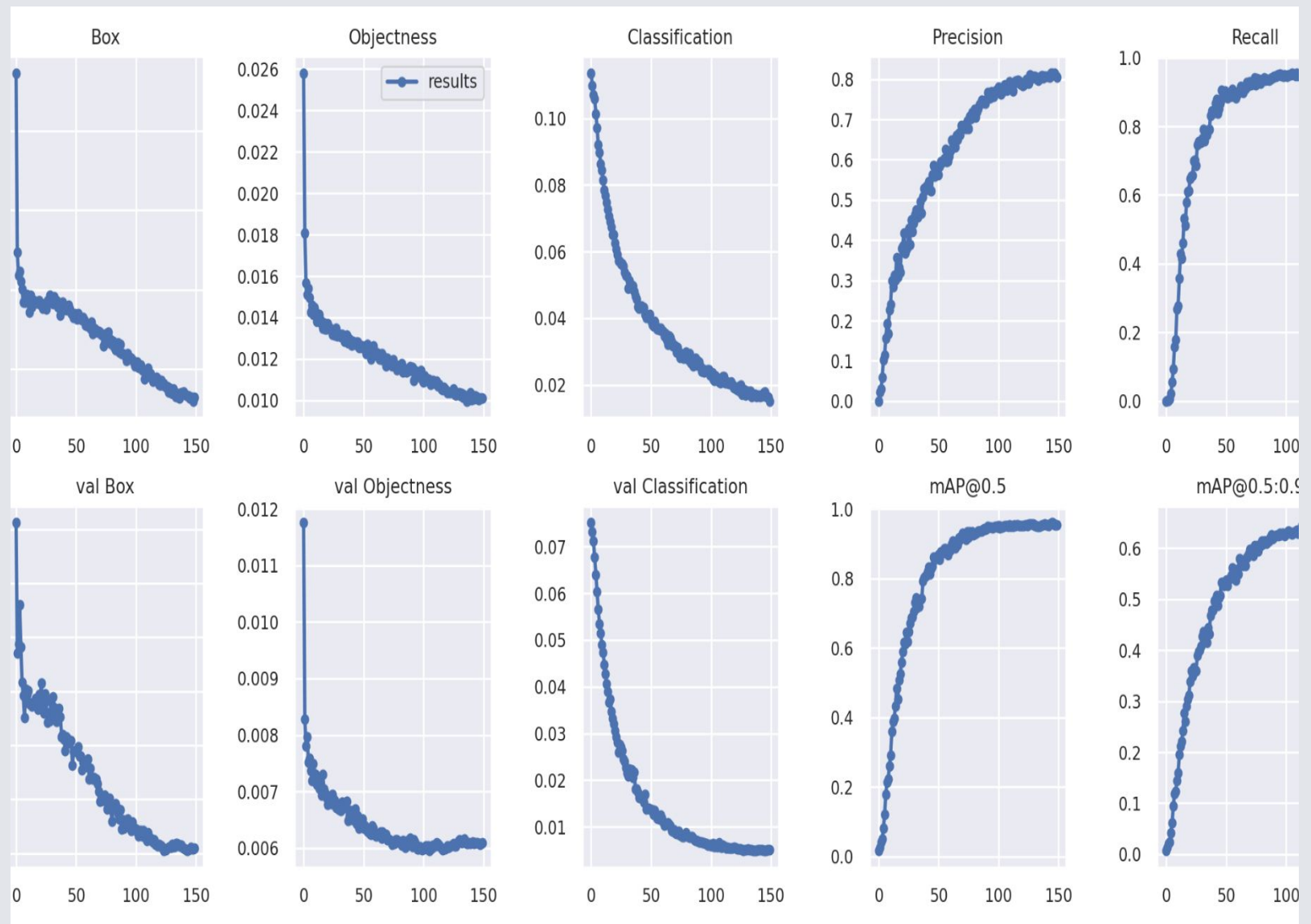


```
.8323274, 'labe
x': 177}], 'dur
.8197195, 'labe
x': 164}], 'dur
edict...
nferences took: 0.6321148872375488 seconds.
'success': True, 'predictions': [{'confidence': 0.8324207, 'labe
```

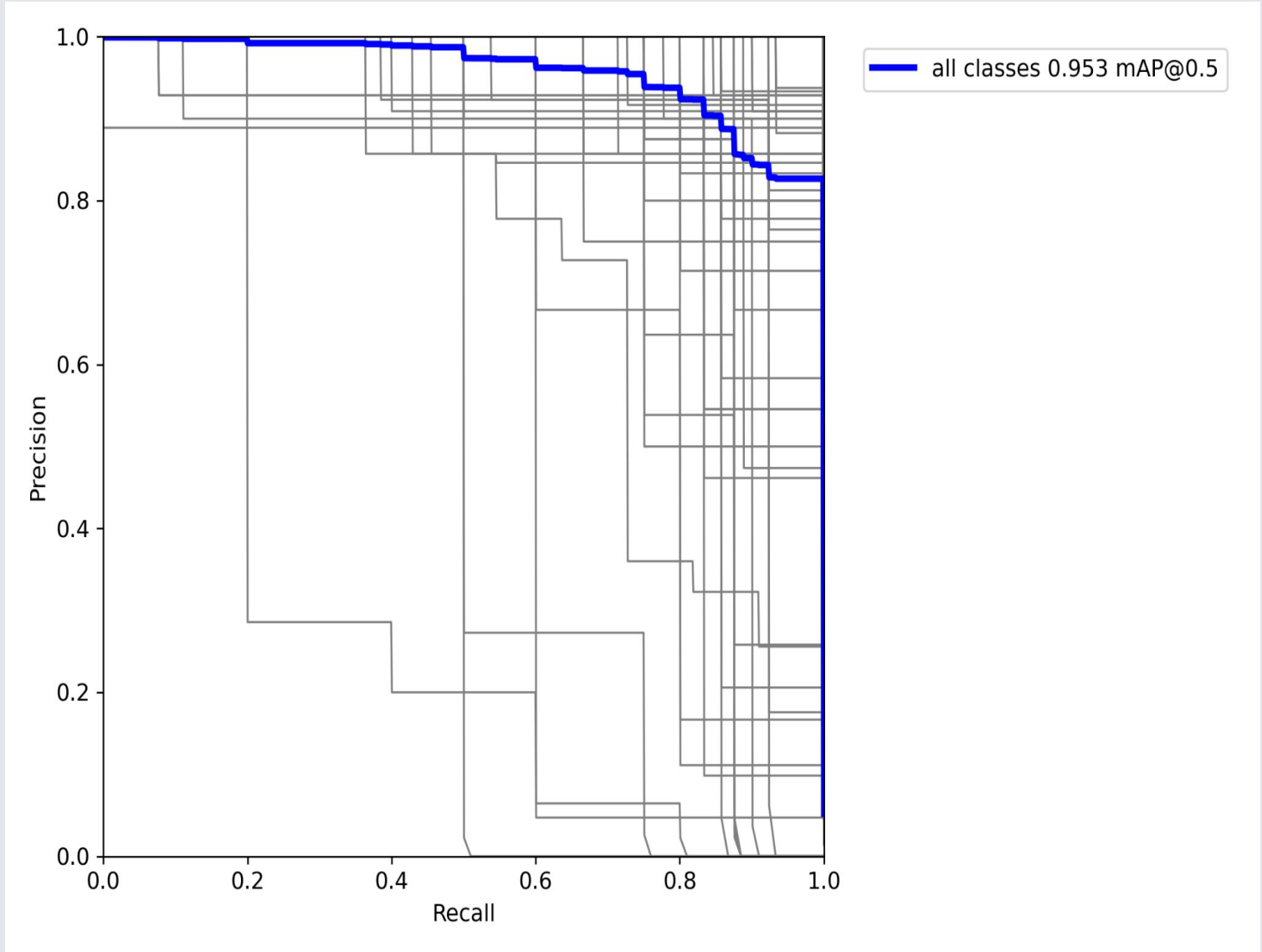
The deployed model performing impressively "in the wild".

A sample batch of test data predicted labels.

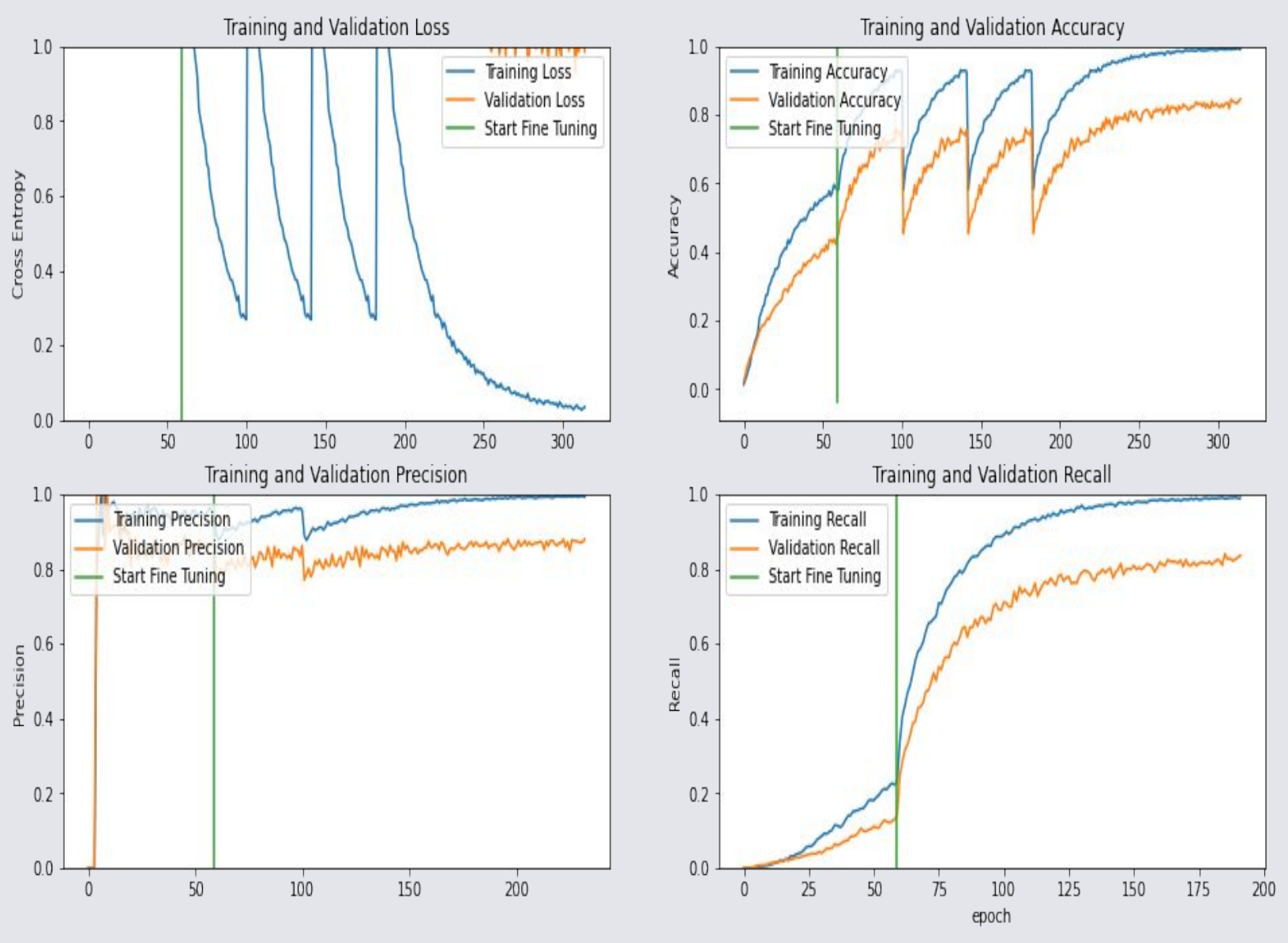
Graphs of YOLOv5's Precision, Recall, mAP of IOU@0.5 and IOU@0.95 as training progressed.



Precision-Recall curve for the YOLOv5 model.



MobileNetv2's Feature extraction vs Fine-tuning metrics/performance comparison.



LIME's explanation of the most important features of an image-level classification predicted by MobileNetv2 as the word "accept".

SCAN ME to go to GitHub (Codebase, Publication link, Demo Video)

What?

An end-to-end and lightweight working prototype, **specifically built for a sub-Saharan African country's sign language** to detect sign language meanings in images/videos and generate equivalent, realistic voice of words communicated by the sign language, in real-time.

Why?

To reduce the communication barrier between the hearing impaired community and the larger society with a focus on sub-Saharan Africa, which is one of the two regions with most cases of hearing disabilities while also being the region with the fewest number of solutions to solve this disconnect.

Lack of solutions for this problem in sub-Saharan Africa is mostly due to two factors:

- The sign language data in the region is low resourced,
- There are increasing complexities and advanced tools required to deploy these solutions in real-life environments.

How?

- Created a novel dataset for a sub-Saharan country sign language (using Nigerian Sign Language as a case study) with over 5000 images across 137 words (incl. 27 alphabets letters). Dataset was created by;
  - a TV sign language broadcaster from OGBC,
  - 20 teachers and students from 2 special education schools in Nigeria.
- Using LabelImg, images were annotated for Object Detection in both TXT and XML formats.
- Data Augmentation - HSV manipulation, Scaling, Cropping, LR-Flipping.
- Object detection models - YOLOv5 and SSD using ResNet50 FPN.
- Classification model using a pretrained model - MobileNetv2;
  - 60 epochs of training using feature extraction only
  - 140 epochs of training after fine-tuning the model

Metrics	YOLO	SSD	Classification
Recall	**0.9512**	0.7075	0.9355
Precision	0.806	0.6414	**0.9063**
mAP:@0.5	0.9533	**0.9535**	N/A
mAP:@0.95	**0.6439**	0.6412	N/A

- Text-to-Speech Conversion with Pyttsx3.
- YOLOModel deployedt using OpenCV, Docker, and DeepStack server.

Thanks to ML Collective for the compute resources grant!