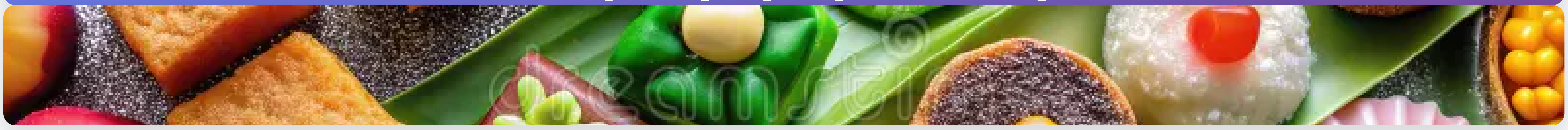
Kuih Classification and Segmentation

Using Ensemble Learning: CNN Segmentation + Vision Transformer

NAIC (AI Technical) Team: CantByteUs

Chin Zhi Xian • Ng Tze Yang • Ong Chong Yao • Terrence Ong Jun Han

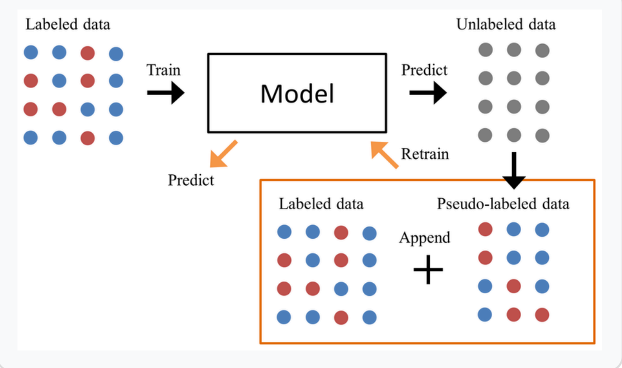


Data Collection & Preparation

- **Scraped ~2500 images/class** from Bing and Google.
- **Removed duplicates** by computing tensor differences between images.
- **Manually filtered** unrelated images and combined with original photos.
- **Annotated original dataset** for segmentation using Label Studio due to a lack of existing segmentation-formatted kuih datasets.
- **Final Dataset:** Plateaued at 98 perfectly annotated images per class, split into 90 for training and 8 for validation.
- **Augmentation:** Used Roboflow to triple the dataset size, excluding hue/color adjustments to preserve color-sensitive features.

Pseudo Labelling Method

- **Efficient Annotation:** Used a semi-supervised technique to create the dataset efficiently.
- **Initial Labelling:** Manually annotated 20-30 complex kuih images per class.
- **Iterative Process:** Trained a small YOLO model to annotate remaining images, followed by manual verification, and retrained a larger model on the combined data.



Model Development

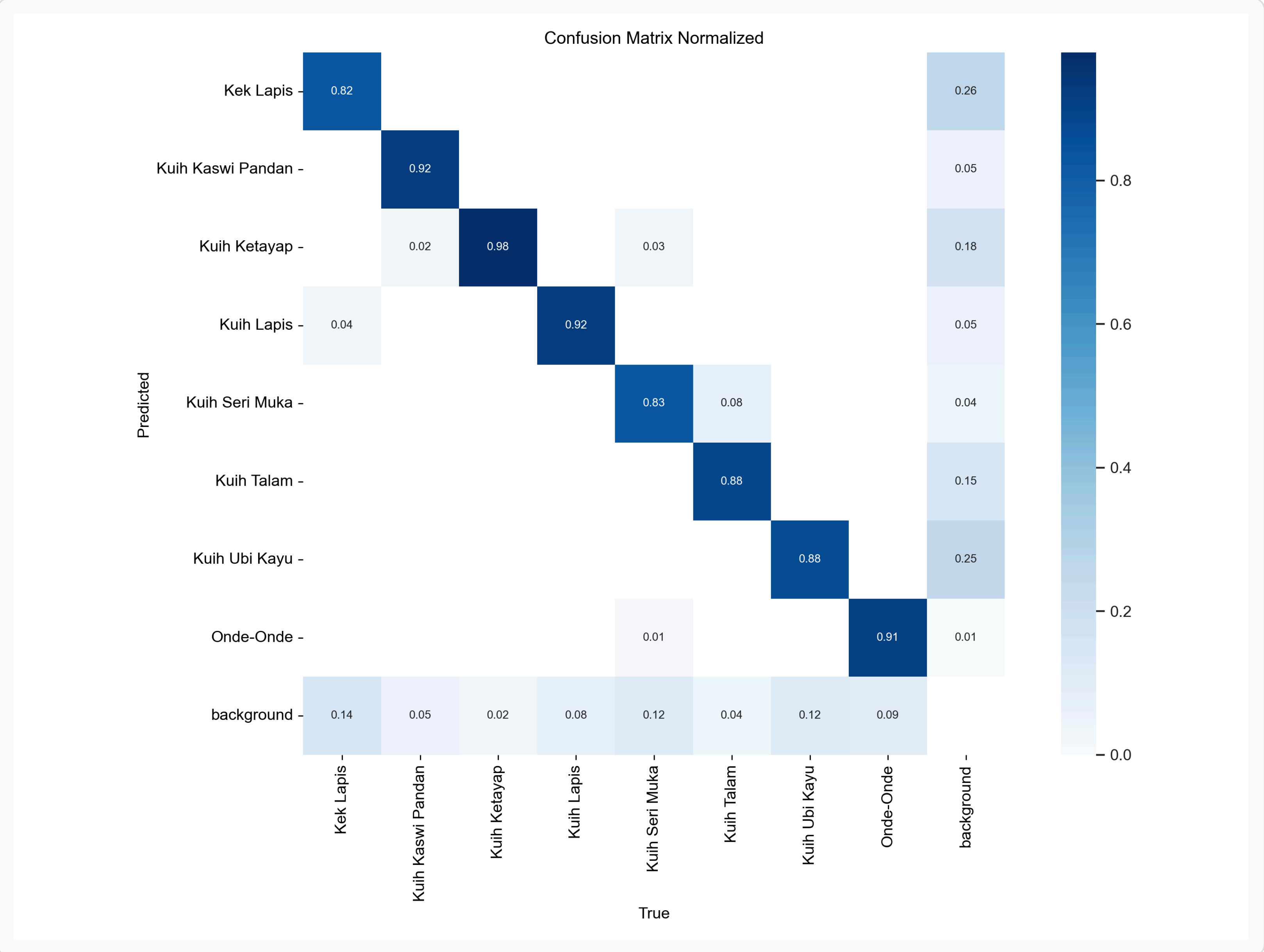
- **Tools:** PyTorch, Ultralytics, and CUDA for GPU acceleration.
- **Initial Challenge:** Large, modified YOLO models tended to overfit on the small, specialized kuih dataset.
- **Solution:** Used a pre-trained **YOLOv11x-seg** model and fine-tuned it on our kuih dataset.
- **Preprocessing:** Normalized image exposure during inference for better model performance.

Why Vision Transformer (ViT)?

- **Global Context:** Splits images into patches and uses self-attention to capture long-distance relationships.
- **Texture Analysis:** Effectively addresses subtle visual similarities (e.g., Kek Lapis vs. Kuih Lapis).
- **Fast Convergence:** Pre-trained on ImageNet 22k, the ViT model converged extremely quickly on our data.
- **Overfitting Risk:** Required careful saving at each epoch to select the best model before it overfit.

Why Segmentation?

- **"Robust segmentation inherently improves classification accuracy."**
- **Focus on the Object:** Prioritizes the kuih itself, significantly reducing background noise and distractions.
- **Impressive Results:** The classification loss plummeted after only a few epochs of training.
- **Near-Perfect Matrix:** Achieved a near-perfect, clean confusion matrix, validating the segmentation-first approach.



Normalized confusion matrix for the YOLOv11x-seg model.

Final Model: An Ensemble Approach

- **Hybrid Power:** Combines the strengths of the **CNN Segmentation model (YOLOv11x-seg)** and the **Vision Transformer (ViT)**.
- **Soft Voting Method:** A soft voting method determines the final classification.
- **Decision Logic:** If models disagree, the class with the highest confidence score is selected.
- **Superior Performance:** This hybrid approach significantly outperformed either solo model, leading to a robust final model.