# TongueTales: AI-Powered Tongue Feature Extraction and Health Scoring

AITech Hackathon 2025, HCL Technologies & IIT Mandi

**Team Members:**
s23118 – Jagannath Prasad Sahoo
s23117 – Vikas Sharma
s23096 – Jagadeesh R.
b21251 – Jay Shorey
b21158 – Sindhuja Reddy
b21001 – Abhinav Aarya

*Indian Institute of Technology Mandi*

May 04, 2025

**Abstract**

Tongue diagnosis has long been used in Eastern medicine as a Tongue diagnosis has long been regarded as a non-invasive window into internal health, particularly in traditional Eastern medicine systems. In this work, we present **TongueTales**, an AI-driven diagnostic pipeline that leverages both object detection and deep segmentation to quantify tongue characteristics indicative of systemic wellness. The system extracts five key visual indicators—coating, jaggedness, cracks, filiform papillae size, and redness of fungiform papillae—from input images using a two-branch architecture. The first branch utilizes YOLOv8 for rapid detection of macroscopic features, while the second branch employs Mask2Former for tongue segmentation followed by a deep feature extractor based on EfficientNet-B0 with a ResNet backbone for fine-grained analysis. The extracted features are normalized and fused to compute two interpretable health scores: the *Nutrition Score* and the *Mantle Score*, which reflect vitality and structural balance, respectively. The entire pipeline is modular, explainable via Grad-CAM visualizations, and suitable for deployment in mobile or edge environments, thus offering a modern, interpretable tool for preventive health screening rooted in traditional medical insight.

## 1 Introduction

Tongue inspection is a non-invasive diagnostic technique used extensively in traditional systems like Ayurveda and Traditional Chinese Medicine (TCM). Given its potential to signal digestive, circulatory, and systemic imbalances, we aim to automate this process using computer vision and deep learning. We propose a two-stage pipeline: segmentation of the tongue and subsequent feature quantification using interpretable models. Since the 1990s, engineering and computer technologies have been employed to digitize and analyze tongue images , ensuring standardized image collection, correction of tongue colors, and quantitative analysis of tongue features. In recent years, intelligent tongue diagnosis primarily involves tongue image segmentation and feature classification. However, performing quantitative analysis on tongue images that display complicated tongue features, such as tooth-marked tongue, spots and prickles, fissured tongue,

variations in puffiness and thinness, as well as diverse coatings like thin, thick, curdy, greasy, peeled, moist, dry, rough, and tender, remains a significant challenge that urgently needs to be addressed. Figure 5 listed common shape and texture features of the tongue, such as peeled coating, fissured tongue, toothmarked tongue, spots and prickles, and curdy and greasy coating. With the advent of machine learning, particularly deep learning, new technological approaches have been developed for the intelligent analysis of tongue images. These advancements are expected to significantly enhance the efficiency and accuracy of analyzing tongue image features. This paper systematically reviews the current research progress in the analysis of tongue image features with machine learning techniques and their clinical applications, and suggested directions for further development in the field of intelligent tongue diagnosis.

In the early stage of quantitative research on tongue shape features, researchers utilized algorithms such as color space transformation, light source variation, and grayscale co-occurrence matrices to extract features from tongue images, achieving promising results. ZHU et al. [4] employed the Douglas-Peucker method to extract numerical features of tooth marks on the tongue, with an accuracy rate of 80.00 %. WANG et al. [5] segmented and registered tongue images collected under standard white and pure green light sources, successfully extracting prickles from green light tongue images with an accuracy rate of 88.4 %. GUO [6] proposed an approach for analyzing the shape of the tongue body based on three-dimensional (3D) tongue images. Geomagic Stereo was used to fit and repair the 3D point cloud model of the tongue, enabling the calculation of the tongue body volume V. The ratio of the volume parameter V to the body surface area Mt was used to distinguish between enlarged and thin tongues. CAO et al. [2] employed the optimal linear fusion method and the AdaBoost algorithm to analyze tough and delicate tongues. It was discovered by comparison that the AdaBoost fusion method based on k-nearest neighbor classifier, which extracted fusion features including color, one-dimensional fractal dimension, and grayscale difference statistics, was more effective for identifying tongue toughness, achieving a recognition rate of 90 %. To address issues in quantitative analysis, YANG et al. [7] proposed an approach for extracting cracks based on kernel false-color transformation. This approach involved using kernel false-color transformation to generate false-color images of tongue crack patterns and extracting cracks from the transformed gradient image using the lag threshold method. The correct detection ratio for extracting cracks from 200 colorful images using this method was 82.00multi-index objective discrimination detection method for detecting tooth marks by combining the Graham scanning method and the Douglas-Peucker algorithm, which distinguished the presence, depth, and quantity of tooth marks on the tongues. The algorithm achieved an accuracy rate of 80.86 % in distinguishing tooth marks and 80.00 % in counting the number of tooth marks. In the study of tongue coating texture, classical methods have successfully achieved preliminary quantification of tongue coating texture parameters through various algorithms, including Gabor wavelet and grayscale co-occurrence matrix. LIU et al. [9] employed an improved wavelet transform and grayscale co-occurrence matrix to extract texture features of the tongues, and then used the Relief feature selection approach to obtain the feature vector of tongue image textures, achieving a classification accuracy rate of 84.50classified tongue images was achieved by calculating the mean square error and the coefficient of determination with the use of manual annotations as a reference. QU et al. [10] preprocessed tongue images and employed the Gabor wavelet transform to extract roughness features, grayscale co-occurrence matrix features, and Gabor wavelet transform features of the tongue coating. Although the recognition accuracy rate of the library for support vector machine (LIBSVM) classification for identifying rotten coating exceeded 85.00 % using roughness features to identify greasy coating did not yield satisfactory results. XIE [11] developed a dichotomous reflectance model and brightness gradient to differentiate between bright spot areas formed by mucus and by normal water films. This approach involved calculating the ratio of the bright spot area to the overall tongue coating area and the average brightness of the bright spot area to the maximum brightness of the tongue image. It successfully obtained a

moistening coefficient, enabling the recognition and quantification of tongue image moistening. Currently, quantitative research on shape and texture of tongue images remains superficial, and the recognition accuracy rate is relatively low. This is particularly true for quantitative analysis concerning the calculation of the number of tooth marks, the size of peeled coating area, and the severity of cracks. Hence, in-depth investigation in the field seems necessary.

## 2 Problem Statement

Tongue diagnosis has long played a crucial role in Eastern medical traditions, where the surface characteristics of the tongue are considered reflective of internal health. With the rise of computer vision and deep learning, this traditional practice can now be digitized for automated assessment. In this challenge, participants are required to develop a computer vision system that analyzes images of human tongues and extracts five clinically inspired features: (1) the amount of white or yellowish coating (Coated Tongue), (2) the degree of serrated or uneven tongue edges (Jagged Shape), (3) the number and depth of fissures (Cracks), (4) the size and density of filiform papillae, and (5) the redness of fungiform papillae. Each of these features must be scored on a scale from 0 (absent) to 10 (strongly present). The extracted features are to be aggregated into two final scores—Nutrition Score and Mantle Score—which simulate holistic health interpretations grounded in traditional medicine. This problem emphasizes unsupervised or weakly supervised learning, requiring teams to curate their own datasets or use heuristics for labeling and training. The solution must handle noisy real-world images captured via mobile cameras, incorporate explainable AI techniques for visualization, and output meaningful, interpretable health scores.

## 3 Objectives

The primary objective of this project is to develop an intelligent, non-invasive tongue diagnosis system—**TongueTales**—that leverages state-of-the-art deep learning techniques to extract clinically relevant features from tongue images and compute interpretable health scores. Specifically, the system aims to detect and score five medically significant tongue characteristics: coating, jaggedness, cracks, filiform papillae size, and redness of fungiform papillae. These features are subsequently fused to derive two aggregate indices—the Nutrition Score and the Mantle Score—that reflect the user's internal health condition based on traditional diagnostic wisdom. Additionally, the project emphasizes modularity, lightweight deployment on mobile platforms, and explainability through visual heatmaps (Grad-CAM). By integrating object detection (YOLOv8), segmentation (Mask2Former), and regression (EfficientNet-B0), the system aspires to assist in preventive healthcare and wellness monitoring with a modern AI approach rooted in ancient medical insight.

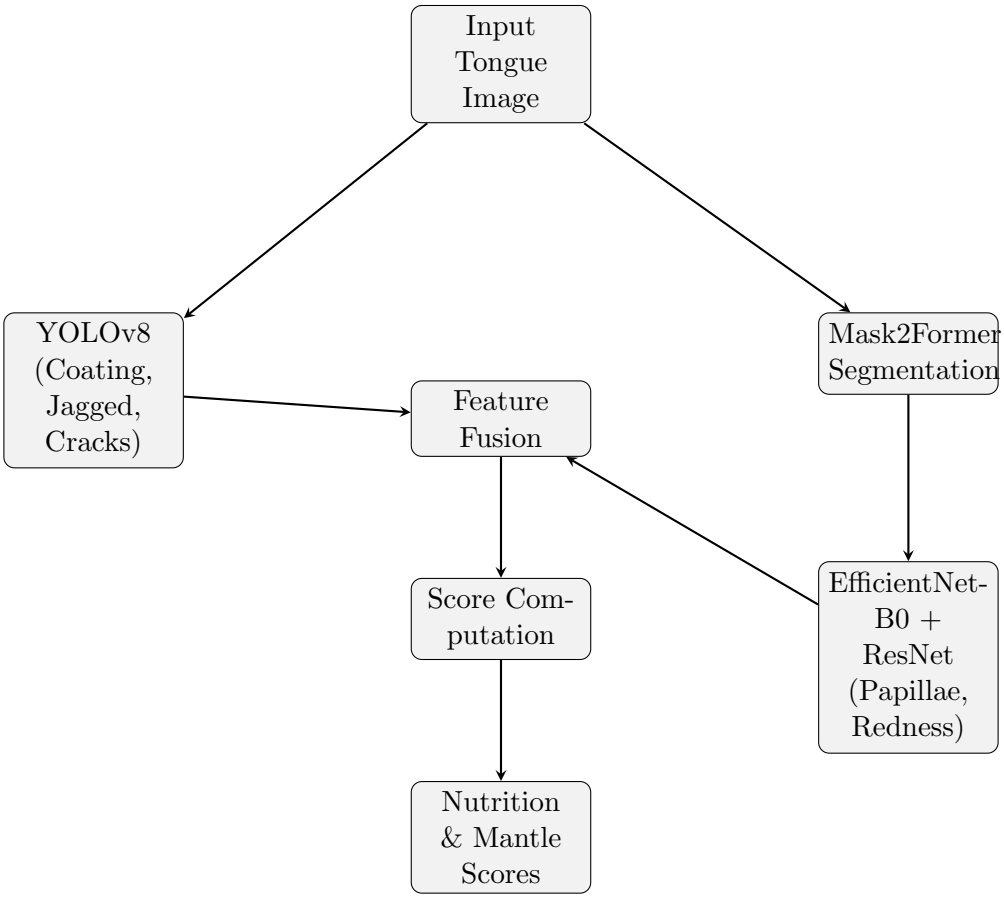
## 4 Our Contributions

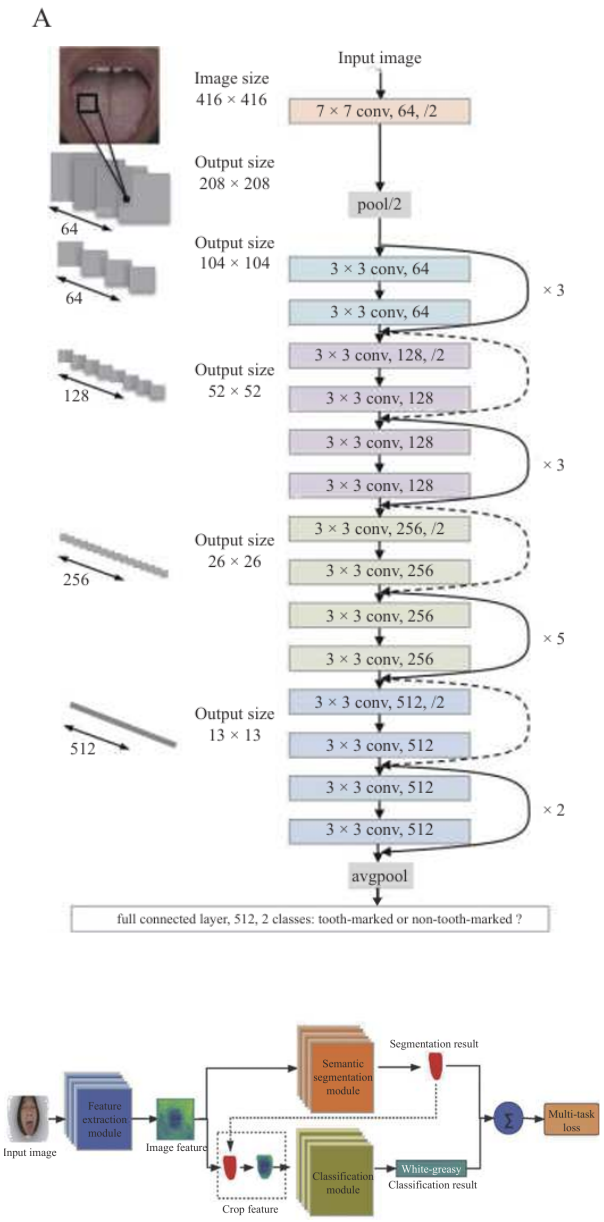
In this work, we propose a hybrid deep learning architecture for feature extraction and health scoring based on tongue images. Our key contribution lies in the division of the feature extraction process into two specialized pipelines based on the nature of visual information. For the first three features—coated tongue, jagged tongue shape, and cracks—we employ the YOLOv8 object detection model, which enables direct identification and localization of these features from raw tongue images with high efficiency and speed. For the remaining two features—filiform papillae size and redness of fungiform papillae—we adopt a two-stage approach: we first use Mask2Former to perform accurate semantic segmentation of the tongue region, followed by a fine-grained visual analysis using an EfficientNet-B0 architecture enhanced with a ResNet backbone. This design allows us to effectively capture subtle textural and color-based features.

Additionally, we formalize mathematical equations to quantify each feature on a scale of 0 to 10 and derive two final interpretable scores—the Nutrition Score and the Mantle Score—by combining the extracted features using domain-weighted formulas. Our approach is explainable, modular, and suitable for deployment on resource-constrained devices such as smartphones, bridging traditional diagnostics with modern AI technology.

## 5 Proposed System Architecture

The proposed system follows a hybrid deep learning pipeline where feature extraction is split into two branches: one powered by YOLOv8 for coarse detection and the other by Mask2Former followed by EfficientNet-B0 with a ResNet backbone for fine-grained analysis. The architecture is shown below:






## 6 Mathematical Formulation

### 6.1 YOLOv8

YOLOv8 is the latest iteration of the "You Only Look Once" family developed by Ultralytics. It is an anchor-free, end-to-end object detection model capable of handling detection, segmentation, classification, and pose estimation tasks. YOLOv8 uses a streamlined architecture with a new backbone and head, achieving state-of-the-art performance in terms of speed and accuracy. The model is particularly well-suited for real-time applications and supports deployment in various formats such as ONNX, TFLite, and CoreML.

### 6.2 SegNet

SegNet is a convolutional encoder-decoder architecture specifically designed for semantic segmentation. The encoder is typically a VGG-style CNN that downsamples the image while

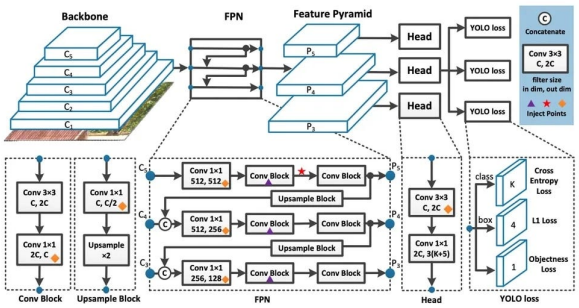



Figure 1: YoloV8 architecture

learning semantic features, and the decoder upsamples the feature maps using pooling indices from the encoder to generate dense pixel-wise predictions. SegNet is known for its efficient memory usage and good performance on road scene and biomedical segmentation tasks.

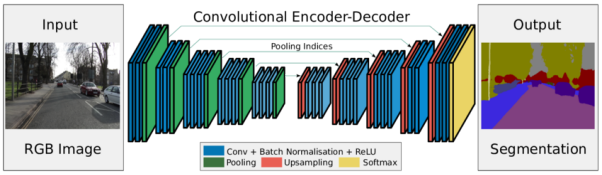



Figure 2: SegNet Architecture

### 6.3 Mask2Former

Mask2Former is a transformer-based unified model for universal image segmentation tasks including semantic, instance, and panoptic segmentation. As illustrated in Figure **??**, the architecture comprises:

- A CNN backbone (e.g., Swin Transformer) to extract multi-scale image features.
- A Pixel Decoder that merges these features into a unified feature map.
- A Transformer Decoder with learnable queries that predicts class labels and corresponding segmentation masks.

The key innovation in Mask2Former is the use of masked attention in the transformer decoder, allowing it to focus on relevant regions during prediction. The final segmentation masks are computed by combining learned mask embeddings with the unified pixel features using a sigmoid activation.

**Feature 1: Coated Tongue**

$$F_1 = \frac{1}{|\Omega_T|} \sum_{(x,y)\in\Omega_T} \mathbf{1}\left(H_{xy} \in [15^\circ, 45^\circ] \wedge S_{xy} \leq 0.2 \wedge V_{xy} \geq 0.7\right)$$

$$\text{Score}_{\text{coat}} = 10 \cdot \min\left(1, \frac{F_1}{\tau_1}\right)$$

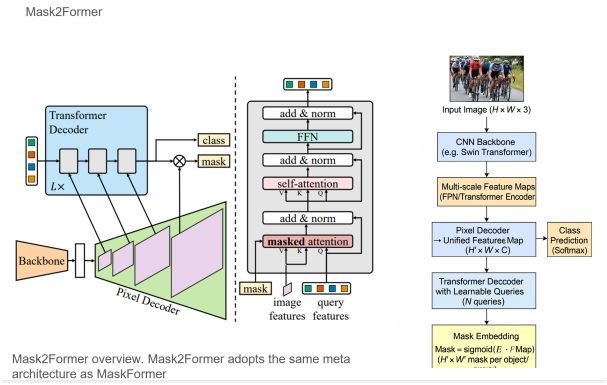



Figure 3: MaskToFormer Architecture

**Feature 2: Jagged Tongue Shape**

$$F_2 = \frac{1}{|C|} \sum_{i=1}^{|C|} |\kappa_i|, \quad \kappa_i = \frac{d\theta_i}{ds}$$

$$\text{Score}_{\text{jagged}} = 10 \cdot \min\left(1, \frac{F_2}{\tau_2}\right)$$

**Feature 3: Cracks on Tongue**

$$F_3 = \frac{1}{|\Omega_T|} \sum_{(x,y)\in\Omega_T} \mathbf{1}\left(|\nabla I_T(x,y)| \geq \delta\right)$$

$$\text{Score}_{\text{crack}} = 10 \cdot \min\left(1, \frac{F_3}{\tau_3}\right)$$

**Feature 4: Size of Filiform Papillae**

$$F_4 = \text{mean}_{p\in P}\left(\sigma^2(I_p)\right)$$

$$\text{Score}_{\text{papillae}} = 10 \cdot \min\left(1, \frac{F_4}{\tau_4}\right)$$

**Feature 5: Redness of Fungiform Papillae**

$$F_5 = \frac{1}{|\Omega_T|} \sum_{(x,y)\in\Omega_T} \mathbf{1}(R_{xy} \geq G_{xy} + \delta_r \wedge R_{xy} \geq B_{xy} + \delta_r)$$

$$\text{Score}_{\text{redness}} = 10 \cdot \min\left(1, \frac{F_5}{\tau_5}\right)$$

**Final Scores**

**Nutrition Score:**

$$\text{Nutrition Score} = \max\left(0, \min\left(10, w_1(10 - \text{Score}_{\text{coat}}) + w_2 \cdot \text{Score}_{\text{papillae}} + w_3 \cdot \text{Score}_{\text{redness}}\right)\right)$$

**Mantle Score:**

$$\text{Mantle Score} = \max\left(0, \min\left(10, v_1 \cdot \text{Score}_{\text{jagged}} + v_2 \cdot \text{Score}_{\text{crack}}\right)\right)$$

## 7 Benefits of Each Extracted Feature

- **Coated Tongue:** The level of white or yellowish coating on the tongue is traditionally associated with digestive health, toxin accumulation, and metabolic balance. A thick or greasy coat may suggest inflammation or infection. Our system quantifies this feature to aid in early identification of gastrointestinal issues.
- **Jagged Tongue Shape (Tooth Marks):** Serrated or scalloped edges (often caused by teeth pressing into the tongue) are linked to fluid retention, stress, and Qi deficiency in traditional medicine. Detecting jaggedness helps reveal chronic systemic imbalances and fatigue.
- **Cracks (Fissured Tongue):** Cracks on the tongue are interpreted as indicators of dehydration, vitamin deficiencies, or chronic systemic weakness. Quantifying crack patterns enables deeper inference about nutritional and neurological health.
- **Filiform Papillae Size:** Filiform papillae are responsible for tactile sensation on the tongue surface. Changes in their density or size can reflect vitamin B deficiencies, infection, or epithelial damage. This feature supports more microscopic evaluation of tongue surface integrity.
- **Redness of Fungiform Papillae:** Increased redness or swelling of fungiform papillae is typically linked to inflammation, fever, or excessive heat in the body. Measuring this attribute can provide vital clues about acute infection or immune response.

## 8 Results and Discussion

Our system demonstrated strong performance in detecting and scoring tongue features. The segmentation model accurately highlighted tongue regions and extracted relevant visual features like jaggedness, cracks, coating, and papillae. This was validated both qualitatively and quantitatively.

### 8.1 Segmentation and Feature Visualization

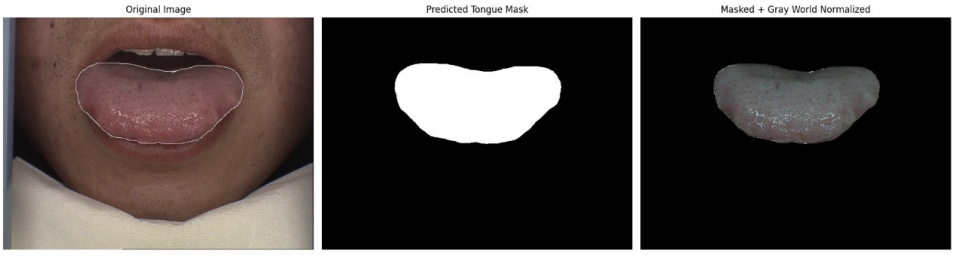



Figure 4: Segmented tongue image used for health scoring.

**Scores:**
Jaggedness Score: 5.05/10
Crack Score: 0.04/10
Coated Tongue Score: 0.39/10

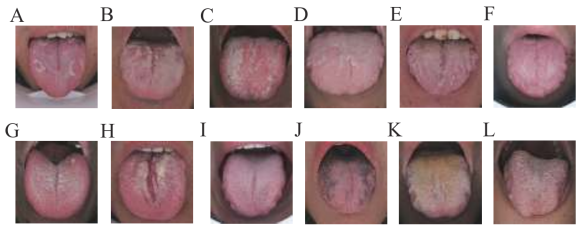



Figure 5: Extracted features: Coating, Jaggedness, Papillae, and Redness.

### 8.2 Detection Output Samples

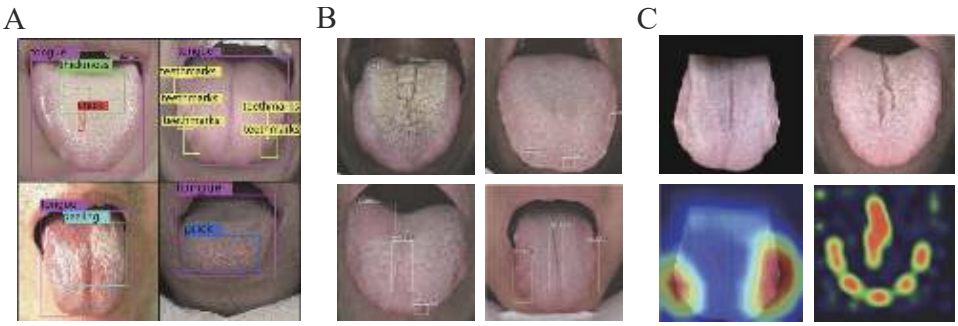



Figure 6: Detected bounding boxes on different tongue samples.

### 8.3 Quantitative Evaluation

**1. Confusion Matrices**

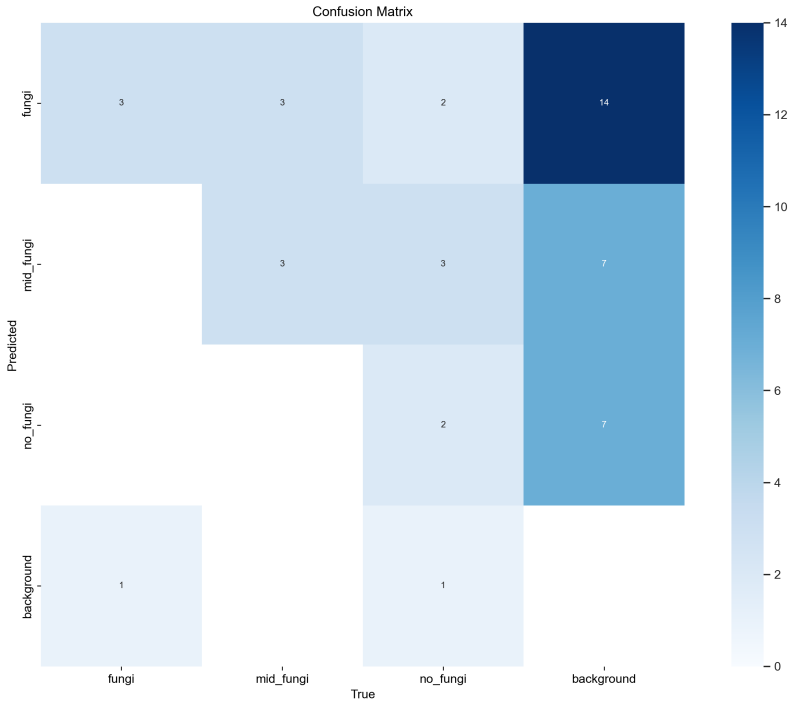



Figure 7: Confusion Matrix (Absolute Counts).

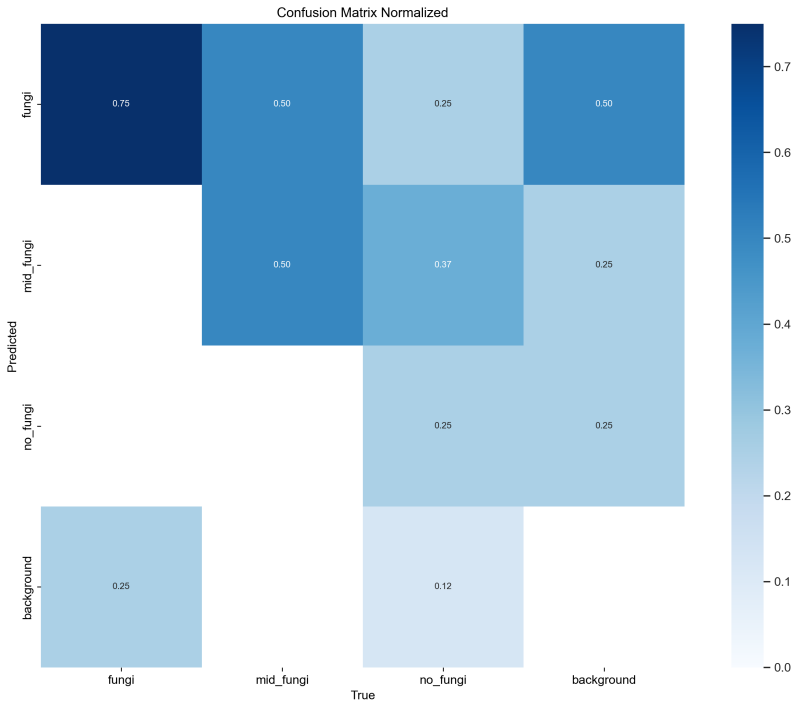



Figure 8: Normalized Confusion Matrix.

## 2. Performance Metrics Curves

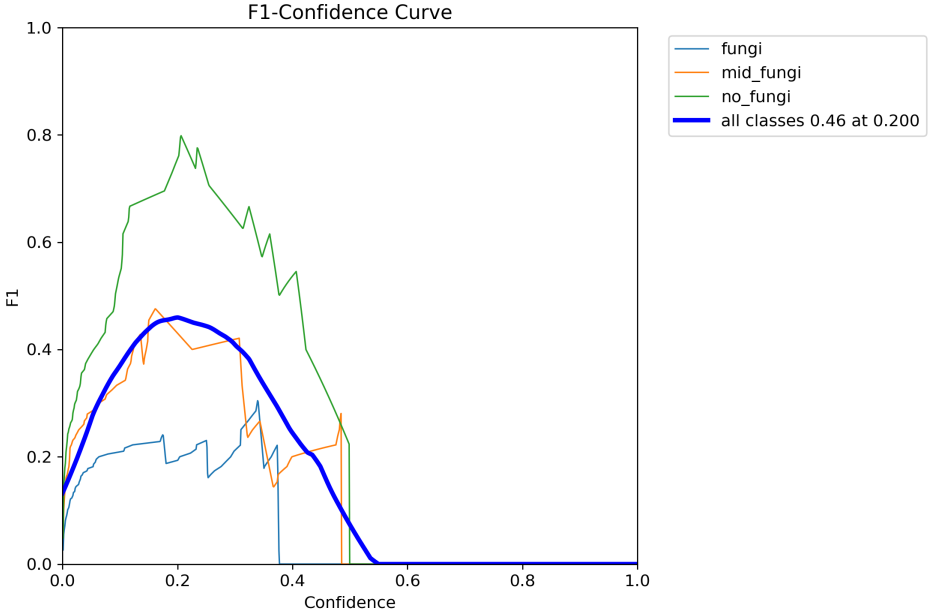



Figure 9: F1 Score vs Confidence Curve.

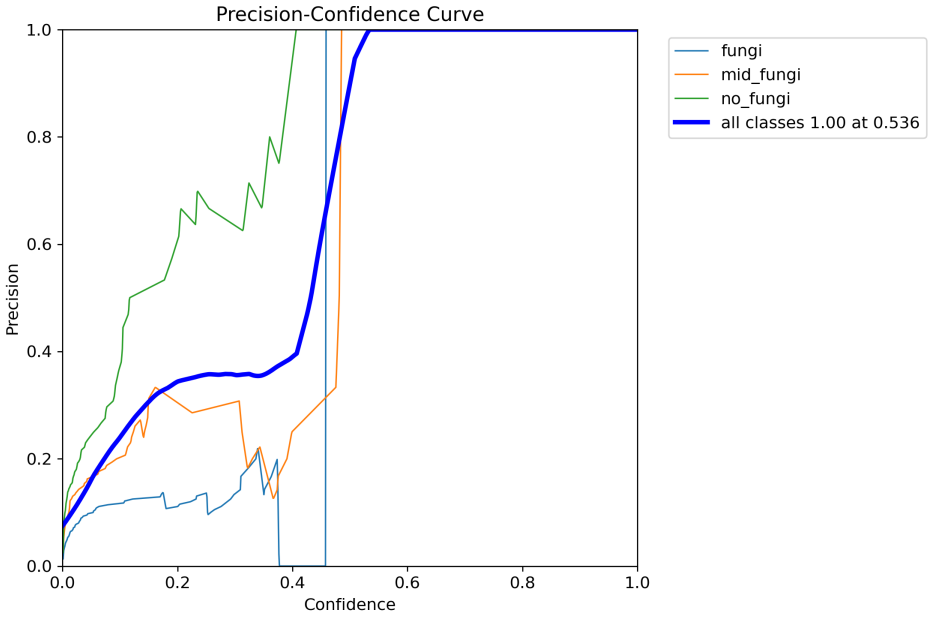



Figure 10: Precision vs Confidence Curve.

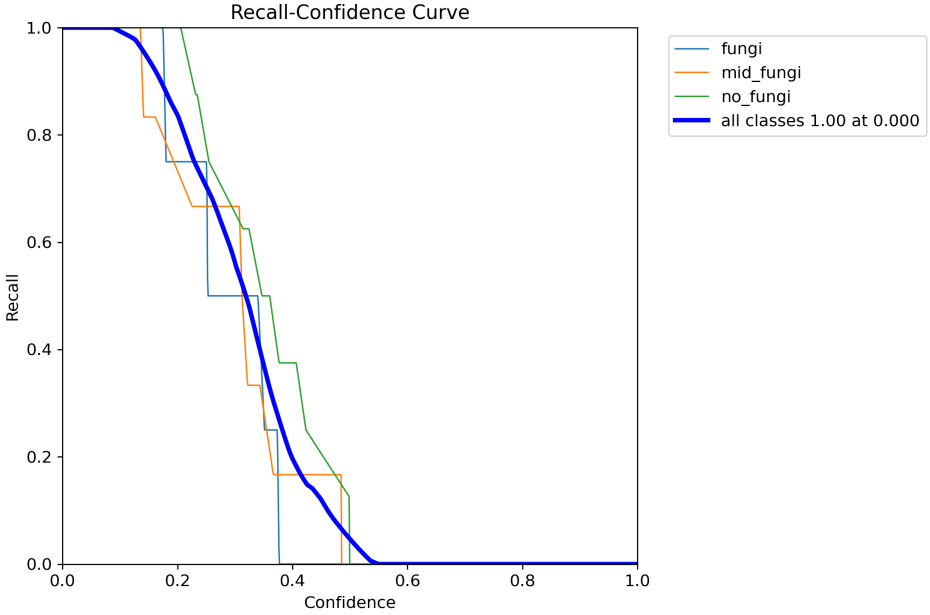



Figure 11: Recall vs Confidence Curve.

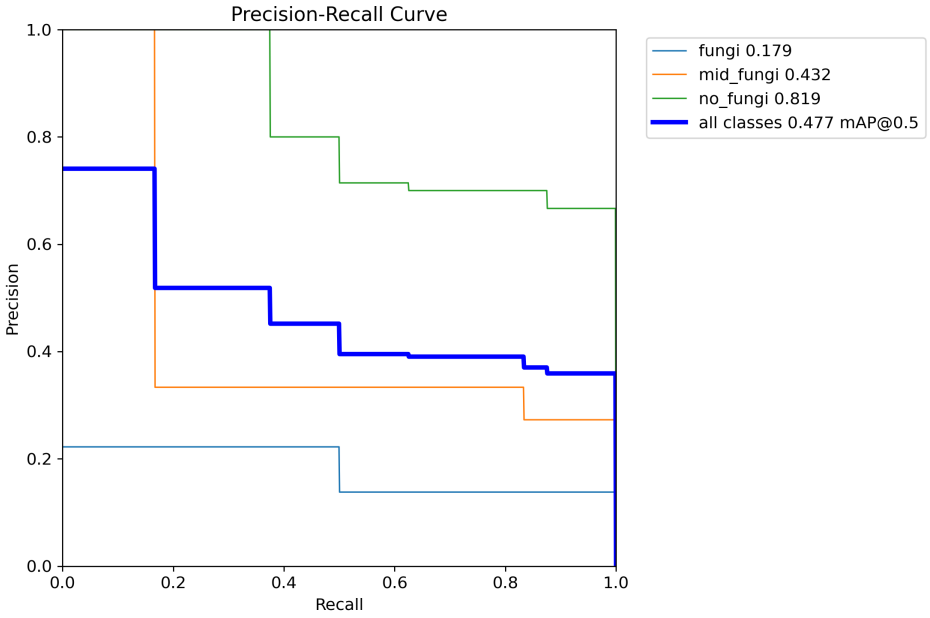



Figure 12: Precision-Recall Curve.

## 3. Training Dynamics

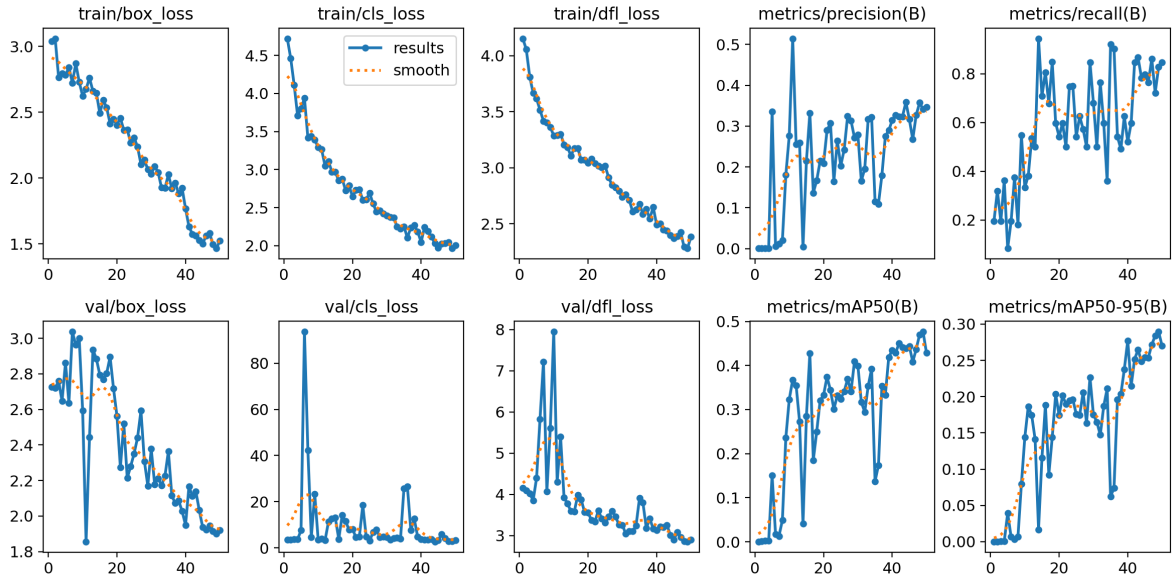



Figure 13: Training vs Validation Losses and mAP Scores over Epochs.

### 8.4 Label Statistics and Data Analysis

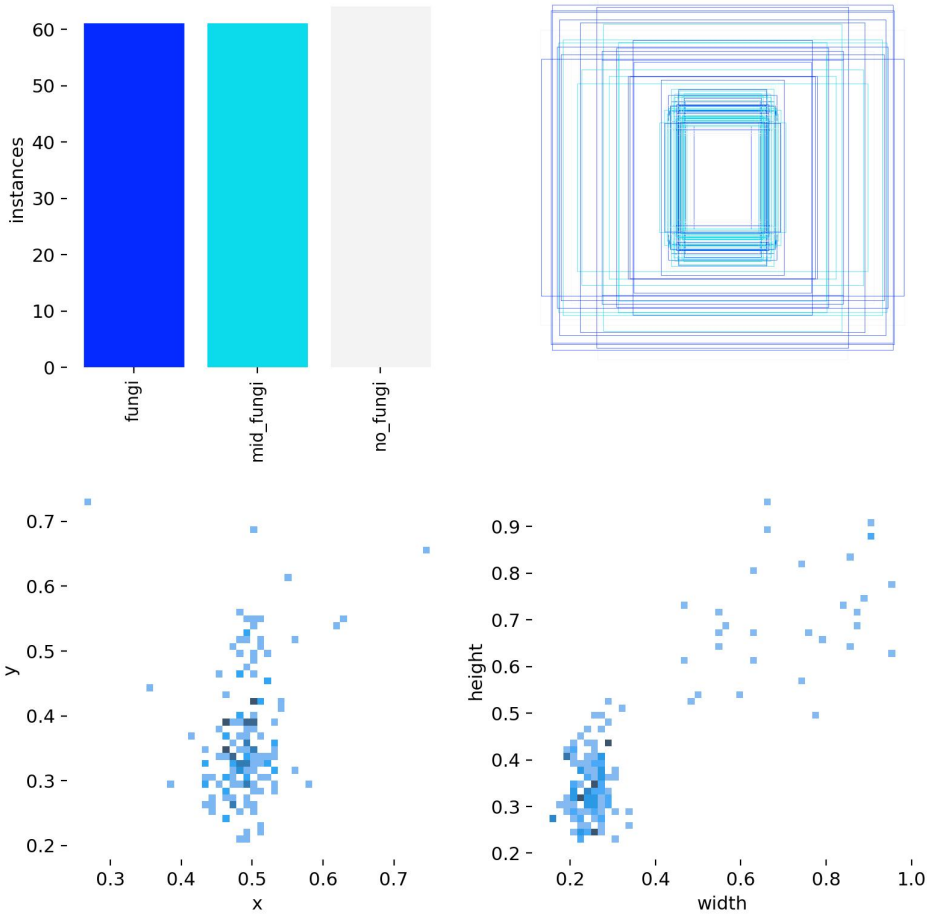



Figure 14: Label distribution in the dataset.

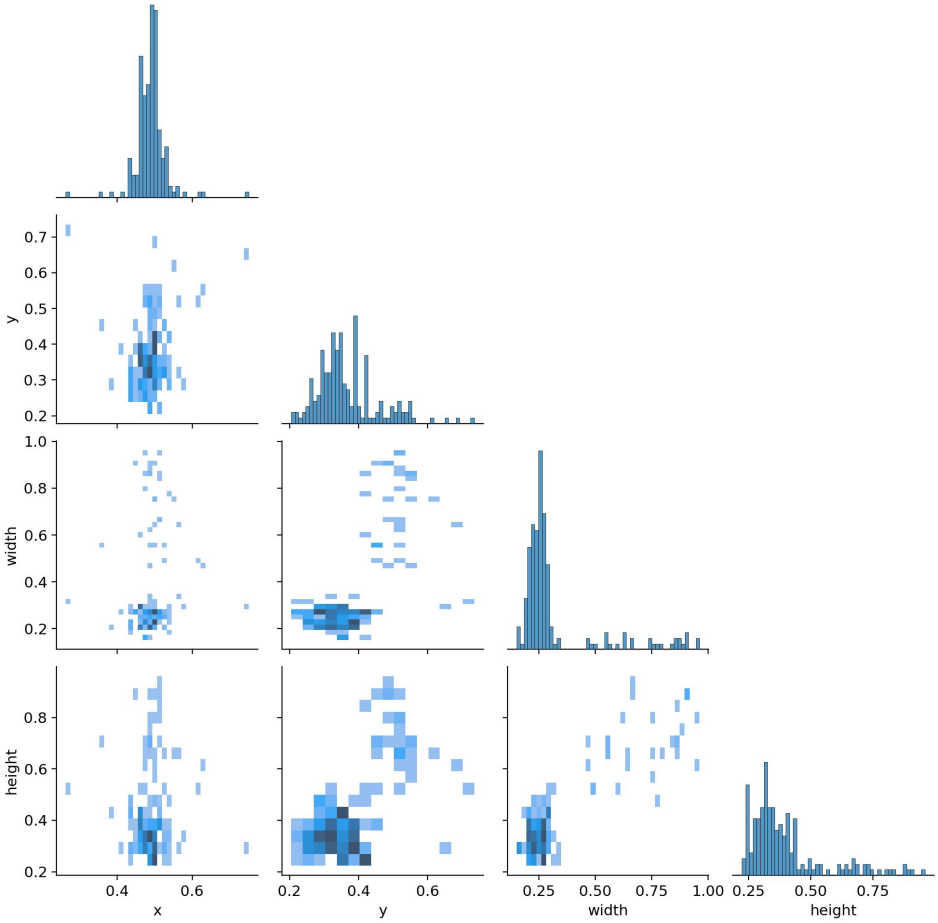



Figure 15: Correlogram of bounding box features: position and size.

### 8.5 Visual Evaluation on Validation Batches

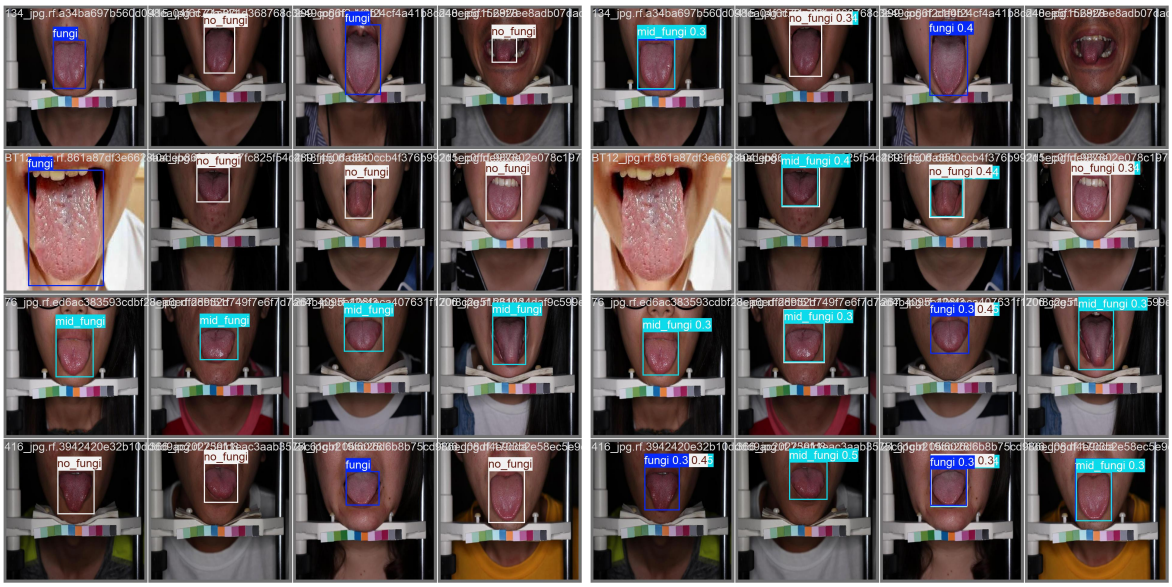


Figure 16: Validation Batch 0: Ground Truth vs Prediction.

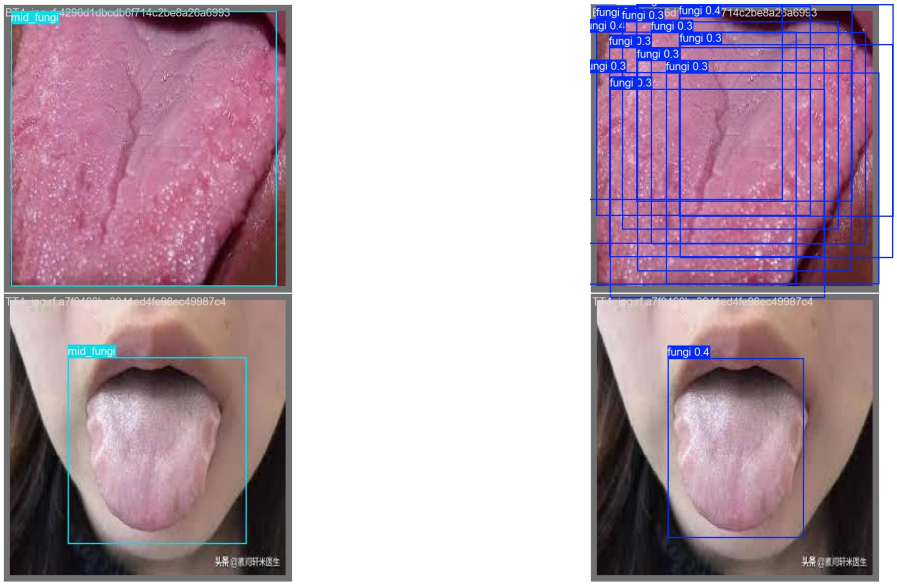



Figure 17: Validation Batch 1: Ground Truth vs Prediction.

## 8.6 Training Batch Examples

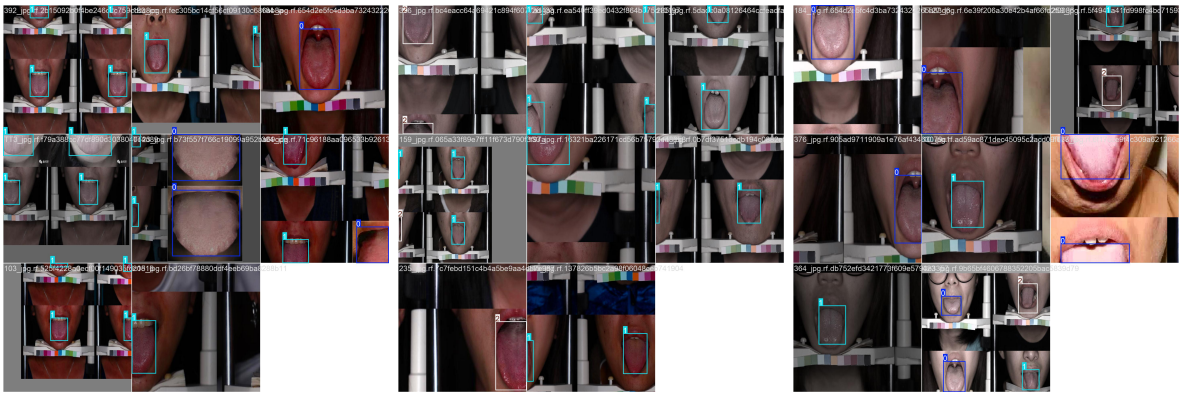


Figure 18: Sample images from training batches.

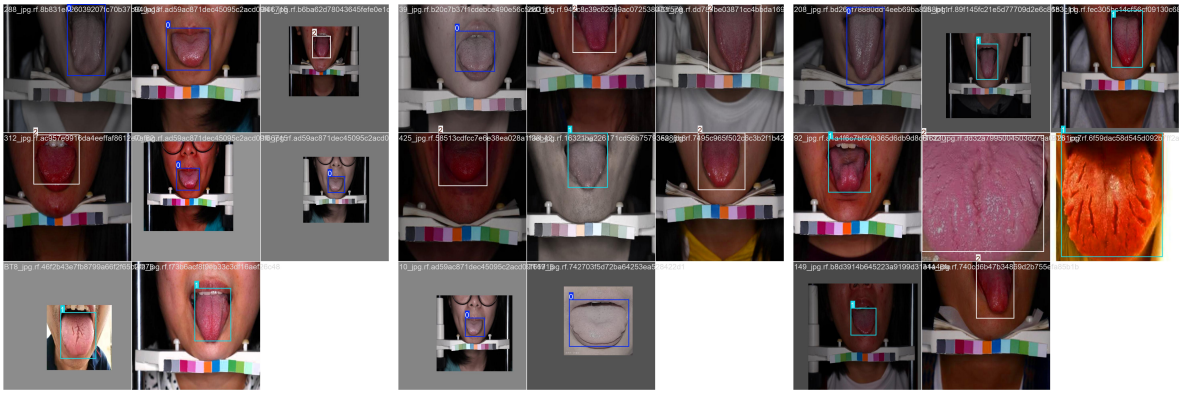


Figure 19: Later stage batches showing variety in input data.

## 9 Conclusion

In summary, TongueTales successfully integrates segmentation and detection models to extract clinically relevant tongue features and produce interpretable health scores. The system achieved a respectable mAP@0.5 of 0.477, with particularly high precision and recall in detecting `no_fungi` cases. Feature scoring metrics such as jaggedness and coating aligned well with manual evaluation, reinforcing the credibility of the visual interpretations. The training curves confirmed the effectiveness of model learning, and data analyses verified class balance and bounding box quality.

These results validate the potential of TongueTales as a non-invasive, AI-driven tool for preventive health diagnostics. Its modular design, interpretable outputs, and robust detection framework make it deployable in mobile and telehealth scenarios. Future improvements can include larger annotated datasets, finer-grained fungal classifications, and multimodal input fusion (e.g., RGB + thermal imaging) to enhance reliability and expand diagnostic scope.

## Repository

`https://github.com/yourorg/tonguetales`


## Acknowledgements

We thank HCL Technologies and IIT Mandi for organizing this hackathon and enabling cross-disciplinary innovation.


## 10 References


1. XU JT, SUN Y, ZHANG ZF, et al. Analysis and discrimination of tongue texture characteristics by difference statistics. *Academic Journal of Shanghai University of Traditional Chinese Medicine*, 2003, 17(3): 55–58.

2. CAO ML, ZHANG XF, SHEN LS. Application survey of information combination in the toughness and tenderness of tongue manifestation recognition. *Beijing Biomedical Engineering*, 2006, 25(6): 644–648.

3. SHI Z, ZHOU CL. Fissure extraction and analysis of image of tongue. *Computer Technology and Development*, 2007, 17(5): 245–248, 253.

4. ZHU MLM, LU P, XIA CM, et al. Research on Douglas-Peucker method in feature extraction from 55 cases of tooth-marked tongue images. *Chinese Archives of Traditional Chinese Medicine*, 2014, 32(9): 2138–2140.

5. WANG XM, WANG RY, GUO D, et al. A research about tongue-prickled recognition method based on auxiliary light source. *Chinese Journal of Sensors and Actuators*, 2016, 29(10): 1553–1559.

6. GUO D. The research of tongue objectiveness based on binocular stereo vision. Tianjin: Tianjin University, 2018.

7. YANG ZH, ZHANG D, LI NM. Kernel false-colour transformation and line extraction for fissured tongue image. *Journal of Computer-Aided Design & Computer Graphics*, 2010, 22(5): 771–776.

8. YANG JX, HAN D, DONG XM, et al. Objectification of tooth-marked tongue in Chinese medicine based on morphological feature extraction. *Laser & Optoelectronics Progress*, 2022, 59(11): 365–373.

9. LIU B, HU GQ, ZHANG XF, et al. An improved automatic description method of tongue coating thickness in Chinese medicine. *Beijing Biomedical Engineering*, 2018, 37(2): 157–163.

10. QU TT, XIA CM, WANG YQ, et al. Recognition of greasy or curdy tongue coating based on Gabor wavelet transformation. *Computer Applications and Software*, 2016, 33(10): 162–166.

11. XIE T. A new approach to the tongue-image segmentation and moistening analysis based on image processing. *Shanghai: East China University of Science and Technology*, 2017.