Breakout Session 4: Track B

Improving AI/ML-Readiness of Data Generated from NIH-Funded Research on Oral Cancer Screening

Dr. Bofan Song
Associate Research Professor, University of Arizona

- **Presentation and Supplement award title:** Improving AI/ML-Readiness of data generated from NIH-funded research on oral cancer screening
- **Funding:** R01DE030682 and R01DE030682-02S1

- **Speaker:** Bofan Song, Associate Research Professor, University of Arizona
- **PI:** Rongguang Liang, Professor, University of Arizona

Contents

- ❑Motivation of the project
- ❑Multi-modal dataset generated from NIH-funded research on oral cancer screening
- ❑Improve AI/ML-Readiness of dataset
- ❑Use of the transformed data in AI applications
- ❑Research outputs
- ❑Future work

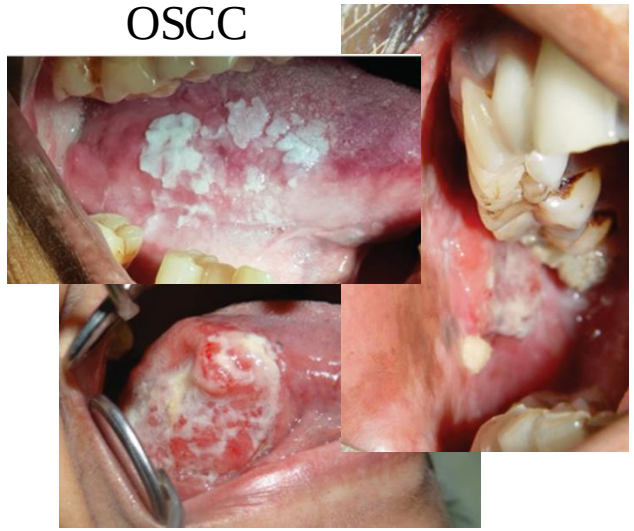
Global Burden of Oral Cancer

- Oral and oropharyngeal squamous cell carcinoma (OSCC) together rank as the sixth most common cancer worldwide, accounting for ~400,000 new cancer cases each year.
  - Two-thirds of these cancers occur in low- and middle-income countries (LMICs), with very high rates in South and South-East Asia.
  - While the 5-year survival rate in the U.S. is 62%, the survival rate is only 20-40% in the developing world.
  - The poor survival rate is mainly due to late diagnosis. Access to cancer prevention, screening, diagnosis, and treatment is a challenge in many LMICs, especially in rural areas with limited health infrastructure.

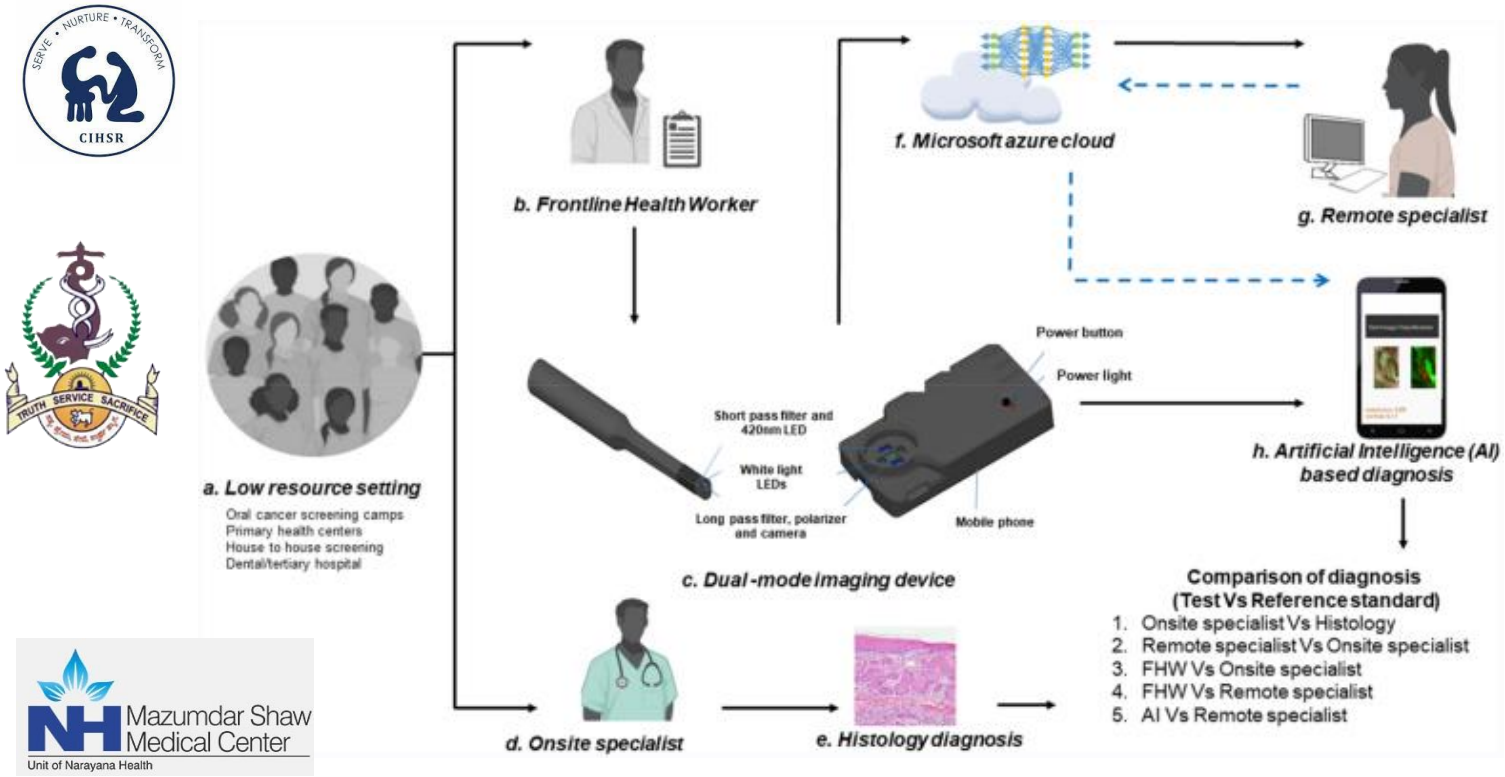
Chewing tobacco

Paan

OSCC

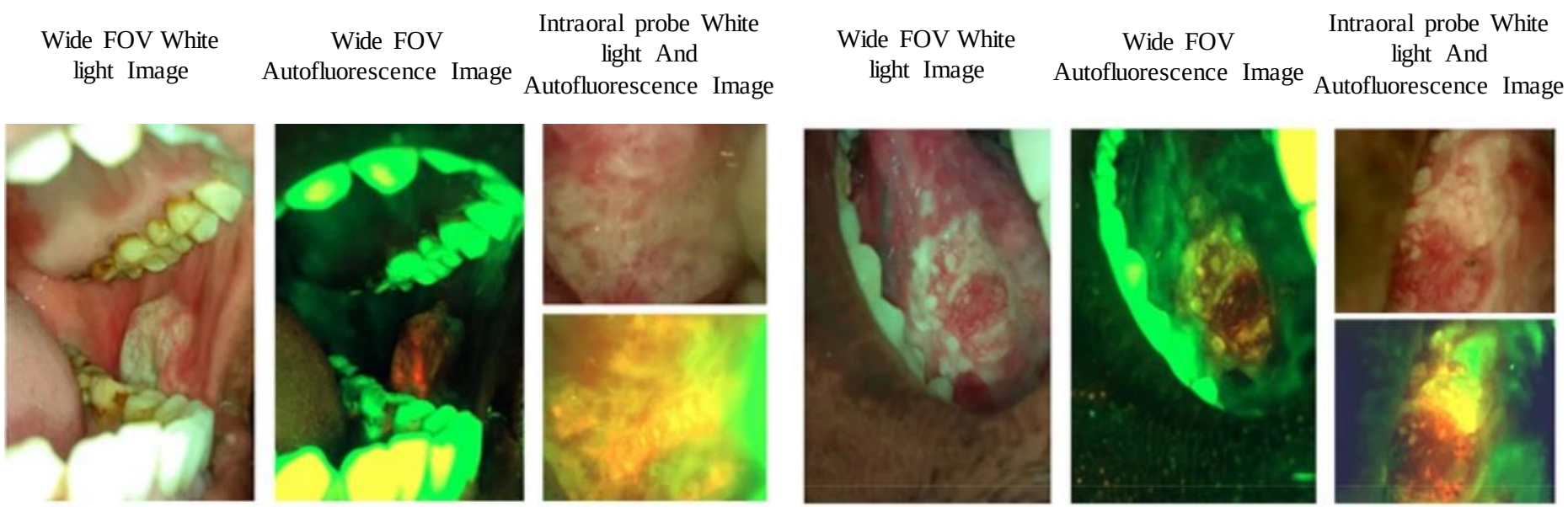


Oral Cancer Screening



- We developed customized multi-modal device for point-of-care oral cancer screening and conduct oral cancer screening using this customized device across multiple clinics in India, reaching thousands of patients within high-risk populations.

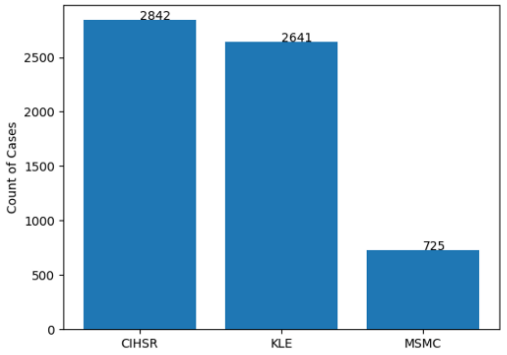
Dataset of oral cancers from high-risk population



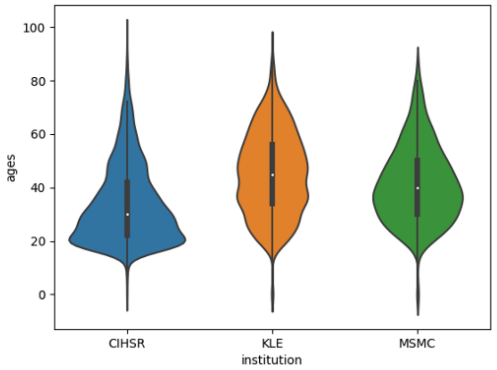
Sex	Age	Employment	Cigarette	Beedi	Tobacco chewing	Tabocco chewing frequency	Arecanut	Arecanut che	Spirit	Lesion site
Female	40	Home-Maker	Never	Never	Never		Current	Daily	Never	Left Cheek
Female	37	Home-Maker	Never	Never	Never		Current	Weekly	Never	Right Cheek
Female	35	Home-Maker	Never	Never	Never		Current	Daily	Never	Left Tongue
Male	50	Service	Never	Never	Current	Daily	Current	Daily	Never	Right Cheek
Female	75	Home-Maker	Never	Never	Current	Daily	Current	Daily	Never	Left Tongue
Male	75	Unemployed	Never	Current	Never		Never		Never	Left Lower Lip
Female	60	Home-Maker	Never	Never	Current	Daily	Current	Daily	Never	Left Tongue
Male	62	Unemployed	Never	Never	Never		Never		Never	Right Lower Lip
Male	25	Service	Never	Never	Current	Daily	Never		Never	Left Cheek

- Our collected dataset contains Wide FOV white light and autofluorescence image, intraoral probe white light and autofluorescence image, as well as corresponding patient information.

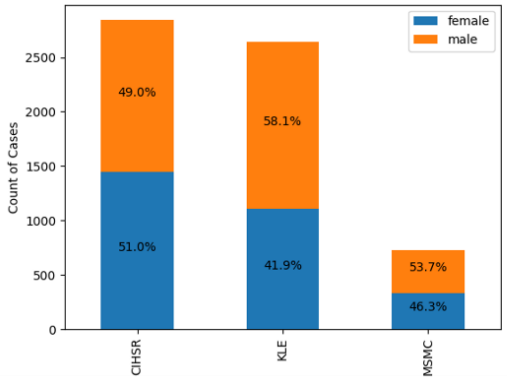
Dataset Demographics



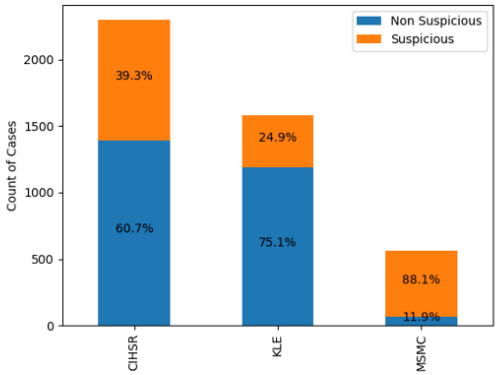
- *Number of patients involved in each institute*



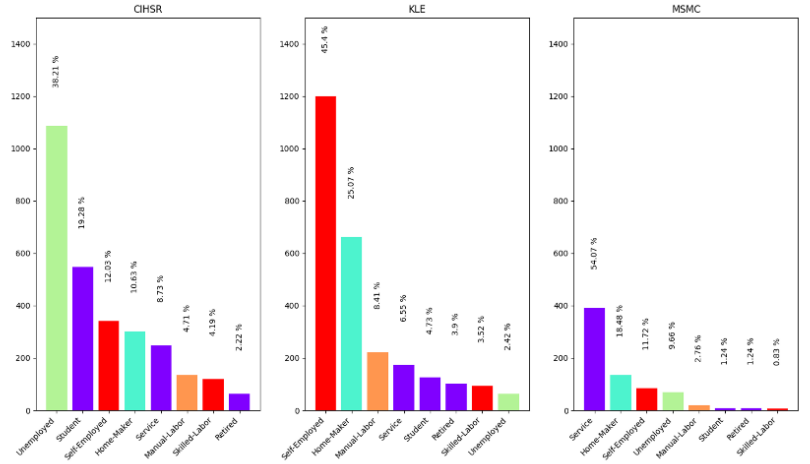
- *Age distribution of patients involved in each institute*



- *Gender distribution of patients involved in each institute*



- *Number of patients that have suspicious lesions involved in each institute*



- *Occupation distribution of patients involved in each institute*

❑ Demographics information of the dataset, such as distribution of the number of subjects, gender, age and etc. across multiple centers.

Improving data compatibility with AI/ML tools

❑ Data compatibility with AI/ML tools

- The dataset includes patient information such as lesion sites, tobacco use history, sex, age, employment status, and other subjective descriptions that are not directly compatible with AI/ML tools like PyTorch or TensorFlow.
- These strings or categorical data types contain essential patient information, which could potentially improve the accuracy and effectiveness of AI/ML algorithms for automatic oral cancer diagnosis.
- Therefore, we have converted the non-numeric features into numeric ones that can be directly used by AI/ML tools.

Sex	Age	Employment	Cigarette	Beedi	Tobacco chewing	Tabocco chewing frequency	Arecanut	Arecanut che	Spirit	Lesion site
Female	40	Home-Maker	Never	Never	Never		Current	Daily	Never	Left Cheek
Female	37	Home-Maker	Never	Never	Never		Current	Weekly	Never	Right Cheek
Female	35	Home-Maker	Never	Never	Never		Current	Daily	Never	Left Tongue
Male	50	Service	Never	Never	Current	Daily	Current	Daily	Never	Right Cheek
Female	75	Home-Maker	Never	Never	Current	Daily	Current	Daily	Never	Left Tongue
Male	75	Unemployed	Never	Current	Never		Never		Never	Left Lower Lip
Female	60	Home-Maker	Never	Never	Current	Daily	Current	Daily	Never	Left Tongue
Male	62	Unemployed	Never	Never	Never		Never		Never	Right Lower Lip
Male	25	Service	Never	Never	Current	Daily	Never		Never	Left Cheek

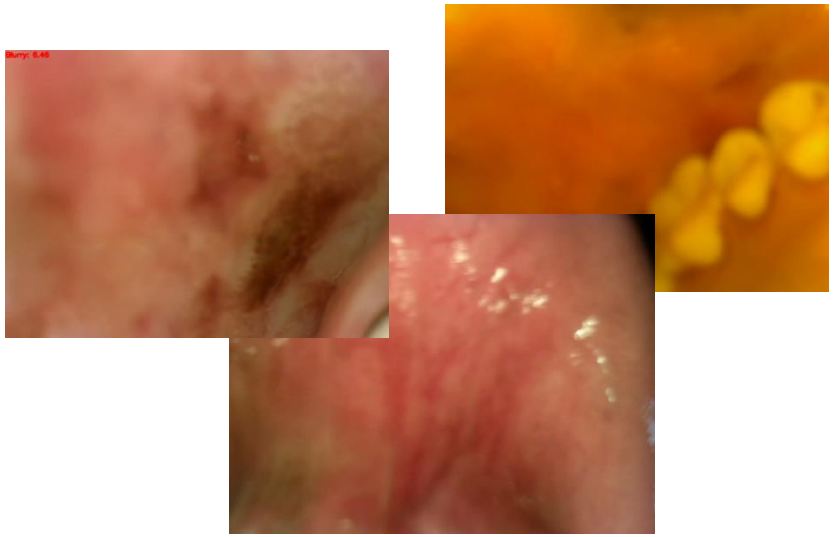
Non-numeric patient clinical information in the dataset

```
age,cig,beedi,tobacco,spirit,beer,la
0111010,000100,000100,000100,000100,
1000100,000100,000100,100000,000100,
0011100,100000,000100,000100,000100,
0101000,100000,000100,000100,000100,
0100010,000010,000100,100000,000010,
0101000,000100,000100,000100,000100,
0111010,000100,100000,000100,100000,
0110111,000100,000001,000100,000100,
0101010,000100,000100,000100,000100,
0100101,001000,000100,100000,010000,
0100101.000001.000100.100000.001000.
```

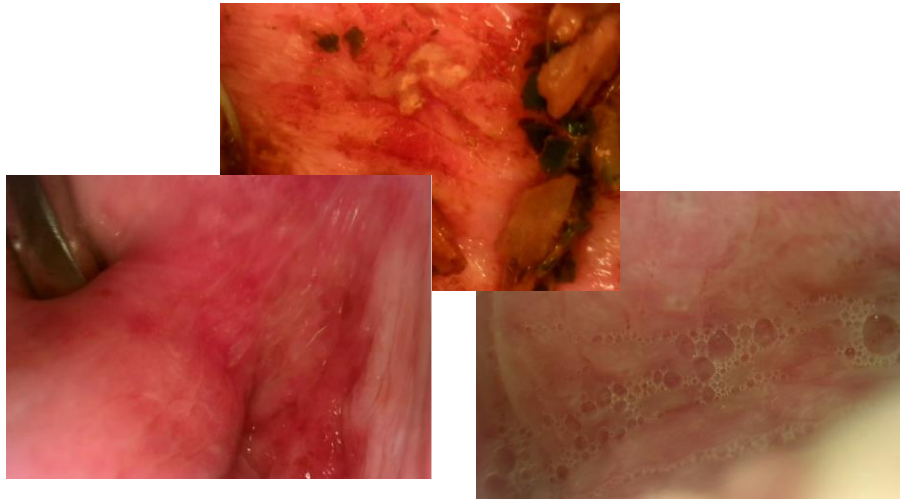
Converted numeric data

Data cleaning

- Identify the blur, low-quality images from the dataset
  - As our dataset was captured during oral cancer screening in low-resource environment using compact device, some blur and low-quality images are inevitable in the dataset.
  - We have identified these instances using the variance of the Laplacian method. Images with lower score typically indicate blurriness. We've flagged these images for researchers' attention.
  - Researchers using this dataset can review the identified blurry images and make a decision about their exclusion before proceeding with AI/ML model development.



Blurry images with lower score



Good images with higher score

Uncertainty in the data labelling

- Estimate uncertainty in the data labelling – label error detection
  - As a large dataset that was labelled manually, erroneous or mislabeled data is inevitable, which can hinder the performance of trained AI/ML models due to labeling inaccuracies.
  - To identify potential label errors within the dataset, we utilized the confidence learning method.
  - Confident learning (CL) is a data-centric approach which focuses on label quality by characterizing and identifying label errors in datasets, based on the principles of pruning noisy data, counting with probabilistic thresholds to estimate noise, and ranking examples to train with confidence.

- After running the confidence learning, the number of potential label error data detected in each modality is showed below.

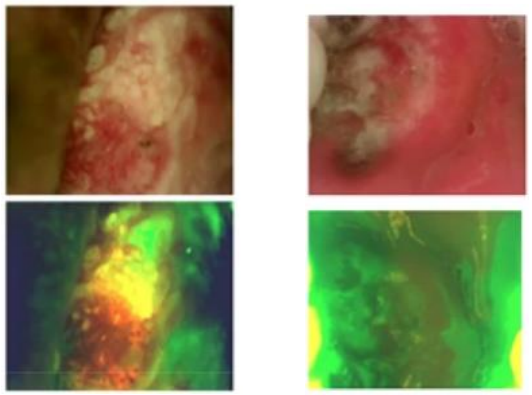
Potential label error images/ All images	Probe autofluorescence image	Probe white light image
Suspicious	272/2897	284/2897
Non-suspicious	488/6069	374/6069

- Challenge: due to the limitation of this algorithm and similarity of oral lesions, distinguishing whether the detected samples were mislabeled or difficult-to-diagnose cases (resulting in the failure of the label error detection method on this dataset) presents a challenge.

- We have reached out to our collaboration specialists to review the identified cases and enhance the overall accuracy of the dataset labeling.

Pixel-level annotation

- Pixel-level annotation for the dataset
  - While an AI/ML model trained with image-level labels may correctly classify an oral cancer image, however, there is a possibility that the model may in fact have its attention focused on an irrelevant region during decision-marking, thereby significantly impacting the model's reliability.
  - Pixel-level annotation offers a more detailed insight into the exact location of lesions. This information could help the researchers utilizing this dataset and aiding AI/ML models in directing their attention to regions of interest during training.
  - Therefore, we are enriching this dataset by providing pixel-level annotations, segmenting the lesion areas within the images.



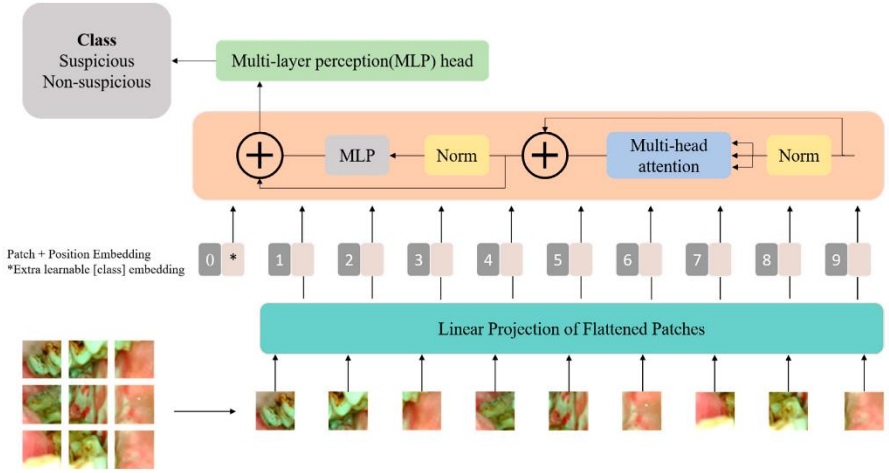
Malignancy Lichen planus

Image-level label

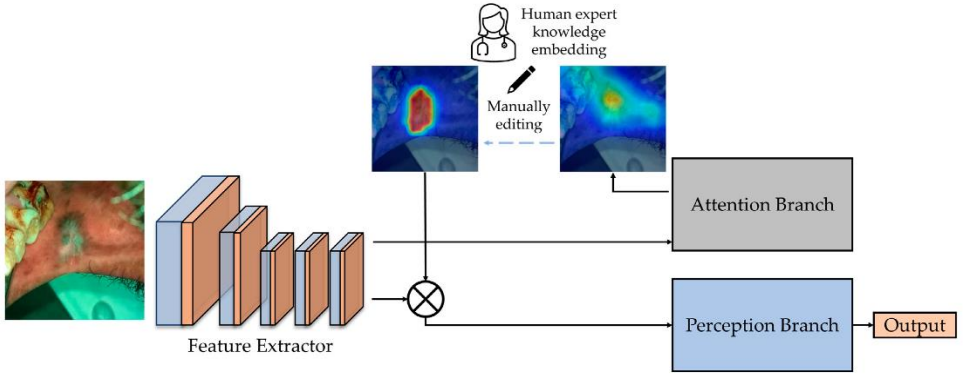
Pixel-level annotation

Use of the transformed data in AI/ML applications

- Use of the transformed data in AI/ML applications
  - Oral cancer image classifier development using vision transformer
  - Improve the model interpretability and reliability using both image-level label and pixel-level annotation



Use the transformed data to develop oral cancer AI classifier



Use the transformed data with both image-level and pixel-level annotation to develop interpretable and reliable oral cancer AI classifier

❑Links to research outputs:

- Song, B., Li, S., Sunny, S., Gurushanth, K., Mendonca, P., Mukhia, N., ... & Liang, R. (2022). Exploring uncertainty measures in convolutional neural network for semantic segmentation of oral cancer images. *Journal of biomedical optics*, *27*(11), 115001-115001. https://doi.org/10.1117/1.JBO.27.11.115001
- Song, B., Zhang, C., Sunny, S., Kc, D. R., Li, S., Gurushanth, K., ... & Liang, R. (2023). Interpretable and Reliable Oral Cancer Classifier with Attention Mechanism and Expert Knowledge Embedding via Attention Map. *Cancers*, *15*(5), 1421. https://doi.org/10.3390/cancers15051421
- Song, B., KC, D. R., Yang, R. Y., Li, S., Zhang, C., & Liang, R. (2024). Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer. *Cancers*, *16*(5), 987. https://doi.org/10.3390/cancers16050987
- Song, B., Dharma, K. C., Li, S., Zhang, C., & Liang, R. (2024, March). Reliable and trustworthy deep learning algorithm design for oral cancer image analysis. In *Design and Quality for Biomedical Technologies XVII* (p. PC1283302). SPIE. https://doi.org/10.1117/12.3007296

Future works

- Future works
  - Oral oncology specialists are reassessing the data identified as potential label errors.
  - Oral oncology specialists are enriching this dataset by providing pixel-level annotations.
  - Multi-class oral cancer AI classification algorithm development using the multi-modal dataset.
    - We're developing a multi-class oral cancer AI classification algorithm using the multi-modal dataset. This involves utilizing different modalities of images and clinical information to enhance classification accuracy.
  - Interpretable and trustworthy AI model development using both image-level labels and pixel-level annotation.