Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

AI-Assisted Triage and Decision Support of Head and Neck Cancer Screening and Diagnosis in Low-Resourced Settings

Min Hun Lee[1] , **Sean Shao Wei Lam**[1,2,3,4] , **Shaun Xin Hong Liew**[1] , **Michael Dorosan**[3] , **Nicholas Graves**[2,8] , **Jonas Karlström**[2,8] , **Hiang Khoon Tan**[2,3,8,9,10] , **Walter Tsong Lee**[5,6,7]
[1]Singapore Management University
[2]Duke-NUS Medical School
[3]Singapore Health Services
[4]Health Services Research Institute
[5]Duke University Health System
[6]Duke Global Health Institute
[7]Duke Cancer Institute
[8]SingHealth Duke NUS Global Health Institute
[9]Singapore General Hospital
[10]National Cancer Centre Singapore

Abstract

The mortality burden of head and neck cancer (HNC) is increasing globally and disproportionately affects people in low-and middle-income countries with limited medical workforce. To address this issue, artificial intelligence (AI) algorithms are increasingly being explored to process medical imaging data, demonstrating competitive performance. However, the clinical adoption of AI remains challenging as clinicians struggle to understand how complex AI works and trust it to use in practice. In addition, AI may not perform well on varying data qualities of endoscopy videos for HNC screening and diagnosis from multiple sites.

In this project, our international and interdisciplinary team will collaborate with clinicians from multiple sites (e.g. Singapore, the U.S., and Bangladesh) to collect a diverse, multi-site dataset. In addition, we aim to design and develop computational techniques and practices to improve collaborations between clinicians and AI for the triage and diagnosis of HNC. Specifically, these techniques include a YOLOv5-based glottis detector, a classifier of patient's status using clinical endoscopy videos, uncertainty quantification techniques, and interactive Vision Language Model-based AI explanations, which will enable clinicians to understand AI outputs and provide their inputs to improve AI. After developing our system, we will evaluate the effectiveness of these computational techniques in enabling AI-assisted point-of-care triage and decision-support for HNC, particularly in resource-limited settings.

1 Introduction

Improving early-stage diagnosis of head and neck cancer (HNC) is important to reducing health burden and patient morbidity [Schutte *et al.*, 2020]. HNC comprises a diverse group of cancers affecting the upper aerodigestive tract, such as oral, pharyngeal, laryngeal, nasal, and salivary gland cancers [Gormley *et al.*, 2022]. These cancers can severely impair essential functions, such as speaking, swallowing, and breathing [Pan *et al.*, 2022]. HNC is the seventh most common cancer globally [Mody *et al.*, 2021] (890,000 new cases and 507,000 deaths annually [Aupérin, 2020]) and has the highest incidences and mortality in developing countries, especially in South and Southeast Asia [Joshi *et al.*, 2014]. Major risk factors of HNC include alcohol consumption, tobacco smoking, and betel chewing especially prevalent in Southeast Asia [Gormley *et al.*, 2022; Pan *et al.*, 2022].

A standard procedure to diagnose HNC in clinical practice is the use of a flexible nasopharyngoscope by trained physicians [Strauss, 2007] to examine and detect abnormalities in the larynx and nasopharyngeal cavity. However, even if early diagnosis of HNC is critical to have better oncological outcomes [Schutte *et al.*, 2020], many low- and middle-income countries lack timely screening and diagnosis of HNC due to their limited medical workforce [Patterson *et al.*, 2020; Ng *et al.*, 2022]. These low- and middle-income countries have an increasing mortality burden of head and neck cancer [Patterson *et al.*, 2020], which results in large and growing economic losses. Specifically, there will be a projected global cumulative loss of $535 US dollars (USD), $180 billion USD losses in Southeast Asia, East Asia, and Oceania, and $133 billion USD loss in South Asia [Patterson *et al.*, 2020]. Addressing this challenge requires urgent efforts to provide timely screening and diagnosic capabilities of HNC, particularly in resource-limited settings, to mitigate mortality and economic impact of HNC.

With recent advances in AI, researchers have demonstrated the potential of AI algorithms to analyze medical data, identify meaningful patterns, and detect diseases with the expert-level competence [Rajpurkar *et al.*, 2022; Mahmood *et al.*, 2021]. These AI algorithms are widely being considered for deployment in clinical practice to improve clinicians' decision-making and patienct care. However, relatively few AI systems have been adopted [Khairat *et al.*, 2018].

A major barrier to adoption is lack of user acceptance and trust [Cai *et al.*, 2019; Lee *et al.*, 2020; Khairat *et al.*, 2018]. As most AI systems function as uninterpretable, "black-boxes" [Cai *et al.*, 2019; Lee *et al.*, 2020], clinicians cannot understand how an AI system has reached its recommendations [Angelov *et al.*, 2021]. Also, clinicians have limited ability to provide inputs or correct an AI system [Lee *et al.*, 2021; Rajpurkar *et al.*, 2022], further reducing their confidence in AI-assisted decision-making.

In addition, AI systems trained on data from a single institute often underperform and fail to generalize across diverse clinical settings [Liu *et al.*, 2020; Rajpurkar *et al.*, 2022]. This results in reduced external validity and underperformance on new patients whose data differ from the training set [Liu *et al.*, 2020; Rajpurkar *et al.*, 2022].

1.1 Problem Statement

This project focuses on the screening and diagnosis of **laryngeal cancer**, a subtype of HNC and one of the most common cancers of the respiratory tract. Laryngeal cancer is among the few oncologic diseases with a declining 5-year survival rate from 66% to 63% [Nocini *et al.*, 2020], underscoring the urgent need for enhanced early detection and intervention.

This project aims to design, develop, and validate computational techniques and practices to improve the collaborations between clinicians and AI for screening and diagnosing laryngeal cancer (Figure 1) in resource-constrained settings. These computational techniques comprise of a classifier of the patient's status using clinical endoscopy videos, uncertainty quantification, and interactive AI explanations for AI-assisted triage and decision support.

For AI-assisted triage, we will leverage an uncertainty quantification module to assist non-specialists identify cases that clinicians should prioritize to review. For AI-assisted decision support, we aim to enhance clinicians' diagnostic capabilities by identifying relevant clinical cases, highlighting important frames and regions of endoscopy images, and providing clinical explanations (e.g. symptoms, and risk factors). In addition, when clinicians review AI outputs and detect errors, they can provide feedback to improve AI. After implementing our AI-assisted system with computational techniques, we will conduct user studies to evaluate its effectiveness.

1.2 Strategy

To achieve the aims of this project, we have formed an international and interdisciplinary team with extensive experience in healthcare services and diverse expertise: medicine (i.e. diagnosis and treatment of HNC), health services research, AI/ML for health, and human-AI interaction. Our team will engage with domain experts, clinicians in Singapore, the U.S., Bangladesh, and the Philippines to iteratively design, develop, and evaluate our AI-assisted triage and diagnosis of laryngeal cancer.

First, we will apply human-centered design approaches to conduct design studies with clinicians, gaining insights into the clinical contexts of laryngeal cancer and identifying how AI can be designed to improve triage and diagnosis of laryngeal cancer. In addition, we will collect endoscopy video datasets from multiple sites (e.g. Singapore, the U.S., Bangladesh, and the Philippines). After developing our proposed AI-assisted triage and diagnosis system, we will conduct user studies to evaluate its effectiveness in improving clinicians' practices of triage and diagnosis for HNC.

1.3 Expected Outcomes & Impact

The proposed research consists of four major activities: (i) data collection and annotation from multiple sites, (ii) formative studies with end users (e.g. clinicians and health professionals) to design the AI-assisted triage and decision support system, (iii) system development using computational techniques, and (iv) system validation with end users.

Our project will establish a robust, real-world database from multiple sites to set a foundation for the development of novel, cost-effective AI-assisted triage and diagnosis capabilities, enhancing the healthcare delivery and the outcomes for patients with head and neck pre-cancer and cancer, particularly in low-and middle-income countries with limited resources. In addition, our project will provide new insights into AI-assisted triage and diagnosis, specifically how clinicians can collaborate with AI and how our system can be deployed to improve triage and diagnosis practices for laryngeal cancer. The computational techniques and findings of our research project will offer valuable knowledge for developing AI-assisted decision-making systems in other domains.

1.4 Alignment with United Nations Sustainable Development Goals (UNSDGs)

Our project contributes to several of the United Nations Sustainable Development Goals (UNSDGs). By developing AI-assisted triage and decision support system for HNC, which has the potential to improve care and outcomes of patients worldwide, our project supports Goal 3 ("Good Health and Well-being: Ensure healthy lives and promote well-being for all at all ages"). In addition, our multinational and collaborative project brings together an interdisciplinary team of researchers and clinicians from institutes in Singapore, the U.S., and Bangladesh. These partnerships support Goal 17: ("Partnerships for the Goals: Strengthen global partnerships for sustainable development), fostering knowledge exchange and enhancing the applicability of our proposed AI-assisted triage and decision support system across diverse healthcare settings. Furthermore, by providing a cost-effective AI-based solution, our project aims to reduce healthcare disparities in resource-limited countries, contributing to Goal 10 ("Reduced Inequalities: Reduce inequality within and among countries").

2 Methods

In this section, we first describe our human-centered design approaches and preliminary dataset, followed by plans for ad-

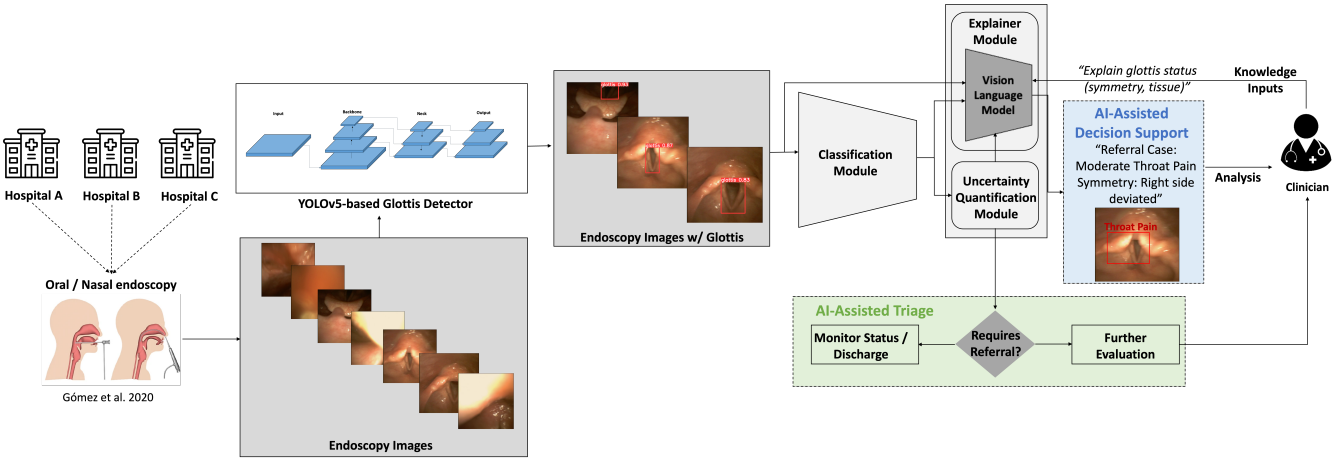


Figure 1: An AI-Assisted Triage and Decision Support System for Head and Neck Cancer Screening and Diagnosis in Low-Resource Settings. Our system first utilizes a YOLO-based glottis detector to select endoscopy frames with glottis. Then, our system classifies the patient's referral status (No referral and non-urgent & urgent referral). For AI-assisted triage, the system utilizes classification outputs along with uncertainty quantification scores to identify cases that clinicians should prioritize for review. In addition, our system leverages a large vision language model to provide interactive explanations of the patient's status to support clinicians' diagnosis of head and neck cancer.

ditional data collection. We then present the technical modules of our research: glottis detector, patient referral status classification, uncertainty quantification, and an interactive vision-language model-based explainer module for AI-assisted triage and diagnosis of laryngeal cancer (Figure 1).

2.1 Human-centered designs, Datasets, & Data Collection

First, our team will conduct formative design studies with clinicians to study and specify clinical context of HNC screening and explore how an AI system can be designed to improve their practices (i.e. screening and decision-making). Drawing on insights from these design studies, we will define the core functionalities of our AI-assisted triage and decision support system for HNC. During system development, we will engage clinicians iteratively to collect their feedback on the system and refine the system's functionalities, ensuring its alignment with real-world clinical needs and workflows.

In this research, we will utilize BAGLS, a publicly available multi-hospital dataset of endoscopy video recordings [Gómez *et al.*, 2020] and additionally collect data from a hospital in the U.S.. We also plan to expand data collection to multiple hospitals in Singapore and low-and middle-income countries (e.g. Bangladesh and the Philippines) to enhance dataset diversity and model generalizability.

BAGLS Dataset

The BAGLS dataset [Gómez *et al.*, 2020] comprises 640 laryngeal endoscopy videos from 380 healthy participants, 210 disordered participants, and 50 with unknown status. The videos are predominately grayscale by having 618 grayscale videos and 22 RGB videos. The endoscopy videos were recorded by multiple clinicians and contain 59,250 endoscopy image frames, each accompanied by a glottis segmentation mask.

Preliminary Dataset & Data Collection Plans

Our team has received ethics approval from the Duke University Health System (DHS) Institutional Review Board (Pro00106209) to access patient data and flexible nasopharyngoscopy videos at Duke Health Centers, which provide care for more than 1,000 HNC cases annually. Our multi-country collaborative research was approved by the DHS Institutional Review Board (Pro00106209) and exempted by the Singapore Health Services Centralized Institutional Review Board (2020/2883).

The preliminary DHS dataset consists of 132 full-color laryngoscopy videos, collected from patients at Duke University Hospital in Durham, North Carolina, USA, between Dec 2019 and Dec 2020. Patients had various medical conditions: laryngeal cancer, benign disease processes, and healthy (no laryngeal pathology). We excluded the patients, who were post-laryngectomy or whose larynx was not visualized on the laryngoscopy video.

The videos vary in lengths, ranging from 5 seconds to 165 seconds. They were recorded under diverse lighting and contrast conditions, orientations, and movements during laryngoscopy procedures, reflecting real-world clinical variability.

A panel of four clinicians (two senior and two junior medical specialists) annotated the DHS dataset. Specifically, videos were classified into three referral levels: Grade 1 (No referral required); Grade 2 (Non-urgent referral or close follow-up in 3-4 weeks); and Grade 3 (Urgent referral).

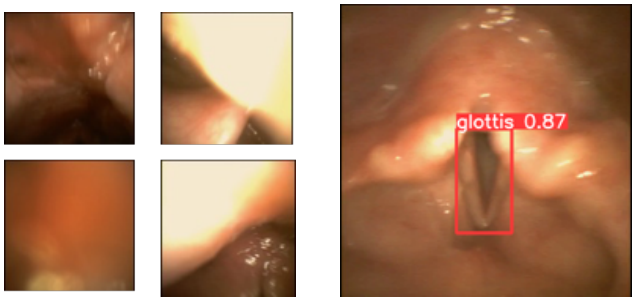
Beyond the preliminary DHS dataset, we plan to expand data collection by acquiring additional laryngoscopy videos and patient data (e.g. clinical reports) from consenting patients at partnering institutions in the U.S. and low- and

middle-income countries (e.g. Bangladesh and the Philippines). This expansion of data collection will enable us to experiment with and enhance generalizability of our system across diverse clinical settings.

2.2 Glottis Detector

Researchers have demonstrated the high performance of AI models in processing medical imaging data and detecting abnormalities [Rajpurkar *et al.*, 2022; Mahmood *et al.*, 2021]. However, real-life data quality often poses challenges, making it difficult to achieve similar high performance in practice. Specifically, we expect to encounter variability in imaging conditions, as different sites may have different imaging devices, protocols, and lighting, resulting in inconsistent image quality [Beede *et al.*, 2020; Wang *et al.*, 2020].

For instance, even if the BAGLS dataset predominantly contains grayscale endoscopy images with a clear view of the glottis, our raw DHS dataset contains RGB endoscopy images that often do not focus solely on the glottis, as the endoscopy videos of the DHS dataset were recorded throughout the entire laryngoscopy procedure (Figure 2a). Developing usable and robust AI algorithms will require addressing these real-world variations and uncertainties to ensure external validity beyond benchmarked datasets.



(a) (b)

Figure 2: RGB endoscopy images during the laryngoscopy procedure (a) without glottis and (b) with glottis.

Previous studies have explored leveraging heterogeneous datasets from multiple sites to extract more robust representations [Liu *et al.*, 2020]. However, these studies have shown only limited performance improvement or even performance degradation compared to models trained on single datasets [Liu *et al.*, 2020].

In this project, we explore image processing and a YOLO-based glottis detector to improve the quality of endoscopy videos. Although clinicians primarily focus on reviewing the vocal glottis area for diagnosing laryngeal cancer, real-world endoscopy videos often contain frames without the glottis throughout the laryngoscopy procedure (Figure 2a). To address this, we will select relevant endoscopy images containing glottis using the YOLO-based glottis detector (Figure 2b), ensuring that the referral classification model is trained on meaningful patient status data.

Initially, we explored to train a UNet model [Li *et al.*, 2018; Gómez *et al.*, 2020] on the BAGLS dataset to segment glottis and identify frames containing it. However, the UNet model, trained on the BAGLS dataset with mostly clean gray scale images, did not perform well on the DHS dataset. Rather than collecting costly glottis segmentation labels for the DHS dataset, we annotated bounding box labels for glottis detection on 2,372 endoscopy images from the DHS dataset using the Roboflow Labelling Service [Dwyer *et al.*, 2022].

To improve glottis detection performance, we applied grayscale augmentation to the annotated dataset in addition to the in-built YOLOv5 augmentations. We then trained a YOLOv5 model to detect the presence of the glottis in endoscopic frames, providing bounding box predictions along with confidence scores (Figure 2b).

2.3 Classification Module

Given endoscopy frames containing glottis, we explore various approach to classify a patient's status into either a non-referral case or a referral case (i.e. non-urgent or urgent referral). These approaches include the Residual Network (ResNet) [He *et al.*, 2016], a convolutional neural network (CNN) model [LeCun *et al.*, 2015], a Long Short-Term Memory (LSTM) model [Hochreiter, 1997]), and a video vision transformer (ViViT) [Arnab *et al.*, 2021].

The ResNet-50 model consists of multiple stages of convolutional layers, followed by (1) batch normalization, a ReLu activation function, and a max-pooling layer, (2-5) residual blocks with two convolutional layers and the identity block, and average pooling and a fully connected layer to generate class probabilities. After training the ResNet50 model on the BAGLS dataset, we utilize an intermediate convolutional layer (i.e. the last convolutional layer of the first stage) of a ResNet-50 model to extract image feature maps from a sequence of endoscopy images (Figure 3).

Using sequential features of endoscopy images from the ResNet50 model, we employ a recurrent neural network (RNN) to classify a patient's status (i.e. non-referral or referral) (Figure 3). Specifically, we utilize a Long Short-Term Memory (LSTM) network, which improves a RNN using memory cells to better capture long-term dependencies [Hochreiter, 1997].

In addition, the emergence of the Transformer architectures, leveraging attention mechanisms, has demonstrated impressive performance in computer vision tasks compared to CNNs [Vaswani, 2017]. To investigate this further, we explore a pure Transformer-based model, Video Vision Transformer (ViViT) [Arnab *et al.*, 2021] for classifying the referral status from endoscopy videos (Figure 3). The ViViT-based classification model first segments an endoscopy video into small clips, extracts spatio-temporal tokens, and encodes them using a series of Transformer layers to capture temporal dependencies across frames [Arnab *et al.*, 2021]. Finally, the output of the temporal Transformer is then passed through an MLP block to predict the patient's referral status.

2.4 Uncertainty Quantification

Even if previous studies have shown AI performance comparable to that of human experts [Rajpurkar *et al.*, 2022; Mahmood *et al.*, 2021], fully automated AI are not ideal for high-stakes domains, such as healthcare. Instead of adopting a fully automated paradigm, this project will explore the

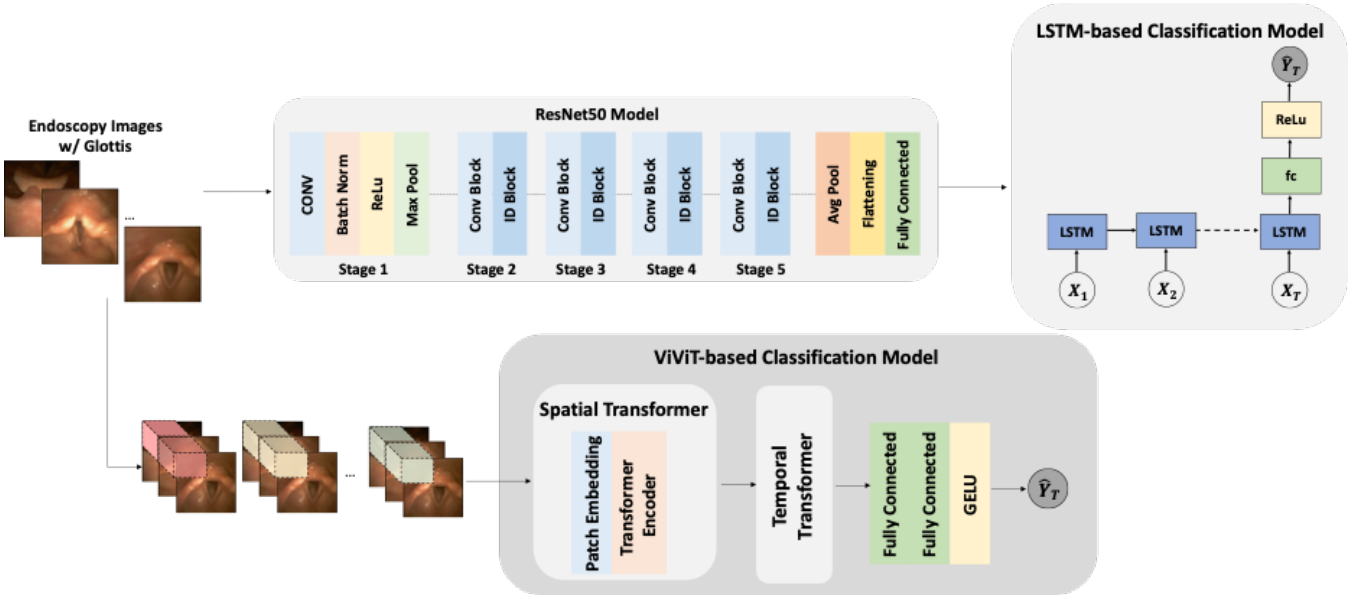


Figure 3: Classification module to determine the patient's status (i.e. no referral or non-urgent/urgent referral). The LSTM-based Model utilizes a ResNet50 model to extract feature maps of endoscopy images. The ViViT-based Model extracts spatiotemporal tokens from the input video and encodes them using transformer layers to handle long video sequences.

emerging paradigm of human-AI collaboration, which emphasizes complementary interaction between clinicians and AI systems [Cai *et al.*, 2019; Lee *et al.*, 2021].

A major impediment to effective human-AI collaboration is the mistrust of AI due to its "perceived" black-box nature [Loftus *et al.*, 2022]. In parallel with the growing emphasis on model explanability [Angelov *et al.*, 2021], there has been increasing research into uncertainty quantification techniques to enhance user trust[Kompa *et al.*, 2021; Loftus *et al.*, 2022]. By incorporating uncertainty scores of AI outputs, users can make more informed decisions and build greater confidence in AI-assisted decision making [Kompa *et al.*, 2021; Loftus *et al.*, 2022].

In this project, we will explore uncertainty quantification (UQ) techniques and describe how uncertainty scores can be computed for effective AI-assisted triage and decision support for HNC screening and diagnosis. Although there is no consensus on the optimal approach to UQ, a common approach involves utilizing the predicted class probability as a proxy for model confidence. However, these predicted class probabilities are not well-calibrated and insufficient for reliabily identifying potential AI model failures [Hendrycks and Gimpel, 2016].

Building upon the work of [Corbière *et al.*, 2019], we will investigate an approach to estimate a confidence score. Given a classification module (Section 2.3), we will train another neural network, called a confidence network to estimate the confidence score of a classification model that is as close as possible to a true class probability. In addition, we will explore alternative UQ techniques, such as Bayesian or Ensemble-based methods [Gawlikowski *et al.*, 2023; Lee and Tok, 2025].

2.5 Interactive Vision Language Model-based Explainer

Although previous studies have shown the potential of clinical decision support systems to provide new insights into clinicians' tasks and improve their decision-making [Rajpurkar *et al.*, 2022], a key impediment to adopting such a system remains the lack of user trust and acceptance [Khairat *et al.*, 2018; Cai *et al.*, 2019; Lee *et al.*, 2020]. Clinicians are often reluctant to rely on systems that operate as a "black boxes", in which clinicians cannot follow how the system generates its output without relevant information. To address this issue, researchers have explored techniques to improve the explainability of AI systems [Angelov *et al.*, 2021]. However, most explainable AI techniques focus on exploring technically oriented explanations, such as highlighting image pixels influencing a model's prediction [Selvaraju *et al.*, 2017], which often lack alignment with clinical reasoning or context.

In this project, we will explore interactive vision-language model (VLM)-based explanations for clinicians to gain new insights into their decision-making tasks (e.g. screening and diagnosis of laryngeal cancer). In particular, this system will enhance clinical interpretability and usability by (1) providing clinically meaningful explanations of AI outputs using VLMs, (2) identifying important image frames [Lee and Choy, 2023] of endoscopy videos or sub-regions within frames that align with relevant clinical contexts, (3) identifying cases that exhibit similar clinical patterns or contexts.

Recent advances in large language models (LLMs) and VLMs have generated significant interest in transforming various aspects of clinical practice including diagnosis, patient triaging, and information retrieval from clinical records [Fries *et al.*, 2022; Moor *et al.*, 2023; Jeong *et al.*, 2024]. Building upon these advancements, we leverage VLMs to generate

clinically relevant explanations of AI outputs (e.g. referral classification and its confidence score) in response to clinician prompts. For instance, a clinician may request *"Provide explanations on glottis status (e.g. symmetry, tissue status)"* (Figure 1). Our interactive VLM explainer utilizes its vision encoder to analyze selected image frames with glottis and AI outputs through its cross attention mechanisms [Fries *et al.*, 2022; Moor *et al.*, 2023] and provide structured clinical explanations aligned with the prompt. These prompts are designed to elicit clinical explanations on AI outputs, such as confidence levels, assessment of anatomical symmetry, tissue condition, abnormality descriptions, and recommended clinical actions. By ensuring that explanations align with standard medical reporting formats and terminology, our interactive VLM-based explainer aims to assist both specialists and non-specialists in interpreting AI outputs, facilitating their decision-making for HNC screening and diagnosis.

In addition, our system allows clinicians to provide their inputs to not only guide the generation of AI explanations, but also contribute to refining the training data and improving model performance. For instance, clinicians can select a region of interest and specify a corresponding clinical context that should be presented in search of similar cases by our system [Cai *et al.*, 2019]. In addition, clinicians may specify a new clinical concept [Kim *et al.*, 2018] or relabel incorrect model outputs [Lee *et al.*, 2022]. These interactive AI feedback mechanisms promote effective human-AI collaboration, enhancing both model interpretability and adaptability to real-world clinical needs.

3 Foreseen Case Studies of AI-Assisted Triage & Decision Support

This project focuses on the screening and diagnosis of laryngeal cancer. We will leverage computational techniques (Section 2) and multi-institutional datasets to support two modes of human-AI collaborations.

1. AI-Assisted Triage: AI identifies severe or uncertain cases that clinicians should prioritize for review
2. AI-Assisted Decision Support:
   (i) Provides data-driven, clinically contextualized insights to assist clinicians in screening and diagnosis of laryngeal cancer
   (ii) Enables clinicians to provide feedback that refines and enhances the AI system's performance

3.1 Evaluation Criteria & Preliminary Results

In this section, we present the evaluation criteria and preliminary results of the computational modules for AI-assisted triage and decision support.

Glottis Detection

For the YOLOv5-based glottis detection model, we split 2,372 annotated endoscopy images from the DHS dataset into 1,937 for training and 435 for testing. At a confidence threshold of 50%, the glottis detection model achieved 88.8% accuracy in identifying endoscopy images with glottis (Figure 2b).

Classification Module

We utilized the BAGLS dataset to train the ResNet50 model to extract feature maps from endoscopy images. The dataset is imbalanced, containing 33,950 non-referral (healthy) and 19,300 referral images. For the experiment, we randomly selected 19,300 non-referral images to balance the dataset, and then split it into training (70%, 27,020 images: 13,510 non-referral and 13,510 referral), validation (20%, 7720 images: 3,860 non-referral and 3,860 referral), and test (10%, 3,860 images: 1,930 non-referral and 1,930 referral) sets.

Model Training and Adaptations: We trained the ResNet50 model using both non-sequential and sequential orders of endoscopy videos, applying the Adam optimizer, cross-entropy loss, and a batch size of 100.

When trained on the non-sequentially ordered dataset and evaluated on the test set, the ResNet50 model achieved 50.33% accuracy as well as Area Under the Receiver Operating Characteristics (AUROC), which reflects the trade-off between sensitivity and specificity.

When trained on the sequentially ordered dataset, the ResNet50 model achieved 81.81% accuracy and 76.17% AUROC on the test set. This substantial improvement in both accuracy and AUROC indicates that rather than shuffling, arranging image frames in their original temporal order enables the model to learn temporal patterns of the glottis, enhancing its robustness for classification tasks.

A model trained solely on data from a single site may struggle to generalize to others. After collecting datasets from multiple sites, we will conduct cross-site validation of our classification module. Specifically, we will explore data augmentation and transfer learning strategies to examine how datasets from one site can be leveraged for another to enhance model generalizability across diverse clinical settings.

Towards Video Vision Transformer-based Endoscopy Video Classification: After extracting sequential endoscopy image features using the trained ResNet50 model, we trained a baseline LSTM-based classification model on the BAGLS dataset. Following the same dataset split as used for ResNet50 training, the 530 videos were divied into training (70%), validation (20%), and test (10%) sets. The baseline LSTM-based model achieved 70.37 accuracy and 65.11 AUROC in classifying non-referral and referral patients. We aim to further explore transformer-based, spatio-temporal video classification models (e.g. TimeSformer [Bertasius *et al.*, 2021], ViViT [Arnab *et al.*, 2021]) to enhance classification performance (Figure 3).

Interactive VLM-based Explainer

Our interactive VLM-based explainer must understand complex medical terminology and the context of HNC and datasets to generate clinical explanations of AI outputs that are comprehensible to clinicians. We will investigate the feasibility of leveraging state-of-the-art general-domain VLMs [Meta, 2024] and VLMs pretrained on public sources, such as PubMed [Jeong *et al.*, 2024] to accurately communicate domain-specific terms to clinicians without misinterpretations. In addition, we will collaborate with health professionals to specify prompts for interactive VLM-based explainer, ensuring that our VLM-based explainer generates actionable

insights for clinicians without overwhelming them with technical details. Furthermore, we will explore how to collect and integrate clinicians' inputs to improve the explainer's contextual understanding, use of medical terminology, and ability to generate clinically relevant, understandable explanations.

Uncertainty Quantification, Interactive VLM-based Explainer, AI-Assisted Triage & Decision Support

For the evaluation of uncertainty quantification (UQ), we will assess its effectiveness in supporting AI-assisted triage of head and neck cancer screening. Specifically, we will investigate how well our approach assists clinicians in identifying uncertain and high-priority cases for their review [Lee and Tok, 2025]. Our approach will be compared against state-of-the-art UQ techniques (e.g. Bayesian Neural Network [Yao *et al.*, 2019], and Deep Ensembles [Lakshminarayanan *et al.*, 2017]).

To facilitate clinical integration, we will collaborate with clinicians from multiple hospitals to develop an AI-assisted triage system that fits existing workflows without increasing cognitive or operational burdens. In addition, confidence scores alone often fail to convey the underlying reasons for uncertainty. To address this, we will work with health professionals to explore the use of an interactive VLM-based explainer, enabling them to interpret and iteratively refine model uncertainty over time. In addition, we will conduct user studies with clinicians to evaluate the effectiveness of our proposed interactive VLM-based explainer for AI-assisted diagnosis of head and neck cancer diagnosis. Specifically, we will examine how well clinicians can improve their diagnostic accuracy for laryngeal cancer with AI outputs from the classification model and context-specific explanations, compared to baseline explanations without contextual information.

3.2 Challenges & Risks

In this section, we discuss the challenges of realizing AI-assisted triage and decision support for head and neck cancer screening in resource-limited settings along with our strategies to mitigate them.

First, even though there are public benchmark datasets of endoscopy videos for glottis segmentation [Gómez *et al.*, 2020], such datasets are usually well-curated for a specific purpose (i.e. glottis segmentation) and may not be representative of the real-world nature of real-world healthcare processes. For instance, the BAGLS dataset has mostly unhealthy patients with muscle tension dysphonia, but lacks data on patients with laryngeal cancer. Curating the real-world dataset of laryngeal cancer is essential to develop an AI-assisted triage and decision support system.

To address this, we have initiated data collection from a U.S. hospital and will expand our efforts to collect a dataset from hospitals in South and Southeast Asia regions (e.g. Bangladesh and the Philippines). As we plan to collect data from multiple sites, we will collaborate closely with domain experts (e.g. clinicians and health professionals) to reduce the variability in data collection procedures and annotations. Also, we may encounter noisy or poor-quality samples from the multi-site dataset. While collecting and processing the dataset, we plan to explore and improve our image processing approach with the YOLOv5-based glottis detection model while remaining open to explore newer object detection architectures or alternative methods as needed.

Although AI has advanced and is increasingly considered for health service delivery [Rajpurkar *et al.*, 2022; Mahmood *et al.*, 2021], there is relatively limited evidence of AI algorithms successfully being deploying in the real-world settings with proven improvements in clinical outcomes. This gap between AI development efforts and successful practical adoption underscores the challenges of integrating AI-assisted systems into healthcare practices. To address this, our team will closely collaborate with our clinical partners to iteratively design, develop, and evaluate our proposed system. Based on continuous feedback from target end-users, we will refine system designs and functionalities. In addition, we will conduct a series of user studies to assess the effectiveness of our approach and investigate how to achieve effective AI-assisted triage and decision support of head and neck cancer screening and diagnosis, particularly in low-resource settings.

Our studies will contribute to developing protocols for trustworthy AI-assisted clinical decision-making [Lee *et al.*, 2024], while investigating challenges, such as over-reliance on AI [Lee and Chew, 2023], systematic detection of distribution shifts, and site-specific model calibration over time. These insights are critical to maintain clinician trust and ensure the safe, long-term deployment of AI systems. In addition, we will explore lightweight model architectures and evaluate system performance under varying computational and environmental constraints in resource-limited settings.

Ethical Statement

In this project, we plan to collect the datasets from multiple hospitals and obtain the necessary ethics approval to collect data from the hospitals. We have received ethics approval from the Duke University Health System Institutional Review Board (IRB) and Singapore Health Services Centralized IRB for our preliminary dataset.

All collected data will be anonymized for annotation and experimental purposes. Personal data, such as age or sex, will be used solely for data analysis and will not be disclosed, ensuring participant privacy. As mentioned in the previous section, we will work closely with the end users (e.g. clinicians, health professionals, and patients) to ensure the ethical development and application of AI in practice. In addition, to facilitate ethical and secure data sharing, we will comply with all relevant local and international data protection regulations.

Acknowledgments

We thank all the participants who contributed to the DHS dataset and gratefully acknowledge Ginny Chen for her support in managing this project. This research is supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (MOE-T2EP20223-0007). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References

[Angelov *et al.*, 2021] Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[Aupérin, 2020] Anne Aupérin. Epidemiology of head and neck cancers: an update. *Current opinion in oncology*, 32(3):178–186, 2020.

[Beede *et al.*, 2020] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[Cai *et al.*, 2019] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.

[Corbière *et al.*, 2019] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32, 2019.

[Dwyer *et al.*, 2022] B Dwyer, J Nelson, J Solawetz, et al. Roboflow (version 1.0)[software], 2022.

[Fries *et al.*, 2022] Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35, 2022.

[Gawlikowski *et al.*, 2023] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[Gómez *et al.*, 2020] Pablo Gómez, Andreas M Kist, Patrick Schlegel, David A Berry, Dinesh K Chhetri, Stephan Dürr, Matthias Echternach, Aaron M Johnson, Stefan Kniesburges, Melda Kunduk, et al. Bagls, a multihospital benchmark for automatic glottis segmentation. *Scientific data*, 7(1):186, 2020.

[Gormley *et al.*, 2022] Mark Gormley, Grant Creaney, Andrew Schache, Kate Ingarfield, and David I Conway. Reviewing the epidemiology of head and neck cancer: definitions, trends and risk factors. *British Dental Journal*, 233(9):780–786, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[Hochreiter, 1997] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[Jeong *et al.*, 2024] Daniel P Jeong, Saurabh Garg, Zachary C Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? *arXiv preprint arXiv:2411.04118*, 2024.

[Joshi *et al.*, 2014] Poonam Joshi, Sourav Dutta, Pankaj Chaturvedi, and Sudhir Nair. Head and neck cancers in developing countries. *Rambam Maimonides medical journal*, 5(2), 2014.

[Khairat *et al.*, 2018] Saif Khairat, David Marc, William Crosby, Ali Al Sanousi, et al. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics*, 6(2):e8912, 2018.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2018.

[Kompa *et al.*, 2021] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, 4(1):4, 2021.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553), 2015.

[Lee and Chew, 2023] Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), 2023.

[Lee and Choy, 2023] Min Hun Lee and Yi Jing Choy. Exploring a gradient-based explainable ai technique for time-series data: A case study of assessing stroke rehabilitation

exercises. In *ICLR 2023 Workshop on Time Series Representation Learning for Health*, 2023.

[Lee and Tok, 2025] Min Hun Lee and Martyn Zhe Yu Tok Tok. Towards uncertainty aware task delegation and human-ai collaborative decision-making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025.

[Lee *et al.*, 2020] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.

[Lee *et al.*, 2021] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[Lee *et al.*, 2022] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Towards efficient annotations for a human-ai collaborative, clinical decision support system: A case study on physical stroke rehabilitation assessment. In *27th International Conference on Intelligent User Interfaces*, 2022.

[Lee *et al.*, 2024] Min Hun Lee, Silvana Xin Yi Choo, Shamala D Thilarajah, et al. Improving health professionals' onboarding with ai and xai for trustworthy human-ai collaborative decision making. *arXiv preprint arXiv:2405.16424*, 2024.

[Li *et al.*, 2018] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[Liu *et al.*, 2020] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging*, 39(9), 2020.

[Loftus *et al.*, 2022] Tyler J Loftus, Benjamin Shickel, Matthew M Ruppert, Jeremy A Balch, Tezcan Ozrazgat-Baslanti, Patrick J Tighe, Philip A Efron, William R Hogan, Parisa Rashidi, Gilbert R Upchurch Jr, et al. Uncertainty-aware deep learning in healthcare: a scoping review. *PLOS digital health*, 1(8):e0000085, 2022.

[Mahmood *et al.*, 2021] Hanya Mahmood, Muhammad Shaban, Nasir Rajpoot, and Syed A Khurram. Artificial intelligence-based methods in head and neck cancer diagnosis: an overview. *British journal of cancer*, 124(12):1934–1940, 2021.

[Meta, 2024] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, Sep 2024.

[Mody *et al.*, 2021] Mayur D Mody, James W Rocco, Sue S Yom, Robert I Haddad, and Nabil F Saba. Head and neck cancer. *The Lancet*, 398(10318):2289–2299, 2021.

[Moor *et al.*, 2023] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[Ng *et al.*, 2022] Sin Wi Ng, Sharifah Nur Syamim Syed Mohd Sobri, Rosnah Binti Zain, Thomas George Kallarakkal, Rahmi Amtha, Felix A Wiranata Wong, Jyotsna Rimal, Callum Durward, Chanbora Chea, Ruwan Duminda Jayasinghe, et al. Barriers to early detection and management of oral cancer in the asia pacific region. *Journal of Health Services Research & Policy*, 27(2), 2022.

[Nocini *et al.*, 2020] Riccardo Nocini, Gabriele Molteni, Camilla Mattiuzzi, and Giuseppe Lippi. Updates on larynx cancer epidemiology. *Chinese Journal of Cancer Research*, 32(1):18, 2020.

[Pan *et al.*, 2022] DR Pan, E Juhlin, AN Tran, Q Wei, S Tang, AT Bui, NG Iyer, and WT Lee. A southeast asian collaborative delphi consensus on surveying risk factors for head and neck cancer screening and prevention. *Global surgery (London)*, 8, 2022.

[Patterson *et al.*, 2020] Rolvix H Patterson, Victoria G Fischman, Isaac Wasserman, Jennifer Siu, Mark G Shrime, Johannes J Fagan, Wayne Koch, and Blake C Alkire. Global burden of head and neck cancer: economic consequences, health, and the role of surgery. *Otolaryngology–Head and Neck Surgery*, 162(3):296–303, 2020.

[Rajpurkar *et al.*, 2022] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

[Schutte *et al.*, 2020] Henrieke W Schutte, Floris Heutink, David J Wellenstein, Guido B van den Broek, Frank JA van den Hoogen, Henri AM Marres, Carla ML van Herpen, Johannes HAM Kaanders, Thijs MAW Merkx, and Robert P Takes. Impact of time to diagnosis and treatment in head and neck cancer: a systematic review. *Otolaryngology–Head and Neck Surgery*, 162(4), 2020.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Strauss, 2007] Robert A Strauss. Flexible endoscopic nasopharyngoscopy. *Atlas of the Oral and Maxillofacial Surgery Clinics of North America*, 15(2):111–128, 2007.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wang *et al.*, 2020] Zhao Wang, Quande Liu, and Qi Dou. Contrastive cross-site learning with redesigned net for covid-19 ct classification. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2806–2813, 2020.

[Yao *et al.*, 2019] Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.