Ophthalmol Ther (2021) 10:703–713
https://doi.org/10.1007/s40123-021-00405-7

COMMENTARY

A Proposed Framework for Machine Learning-Aided Triage in Public Specialty Ophthalmology Clinics in Hong Kong

Yalsin Yik Sum Li · Varut Vardhanabhuti · Efstratios Tsougenis ·
Wai Ching Lam · Kendrick Co Shih

Received: July 29, 2021 / Accepted: September 29, 2021 / Published online: October 12, 2021

ABSTRACT

The public specialty ophthalmic clinics in Hong Kong, under the Hospital Authority, receive tens of thousands of referrals each year. Triaging these referrals incurs a significant workload for practitioners and the other clinical duties. It is well-established that Hong Kong is currently facing a shortage of healthcare workers. Thus a more efficient system in triaging will not only free up resources for better use but also improve the satisfaction of both practitioners and patients. Machine learning (ML) has been shown to improve the efficiency of various medical workflows, including triaging, by both reducing the workload and increasing accuracy in some cases. Despite a myriad of studies on medical artificial intelligence, there is no specific framework for a triaging algorithm in ophthalmology clinics. This study proposes a general framework for developing, deploying and evaluating an ML-based triaging algorithm in a clinical setting. Through literature review, this study identifies good practices in various facets of developing such a network and protocols for maintenance and evaluation of the impact concerning clinical utility and external validity out of the laboratory. We hope this framework, albeit not exhaustive, can act as a foundation to accelerate future pilot studies and deployments.

Keywords: Machine learning; Triage; Ophthalmic specialty clinics; Public health care system

Y. Y. S. Li · W. C. Lam · K. C. Shih (✉)
Department of Ophthalmology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 301B Cyberport 4, 100 Cyberport Road, Pokfulam, Hong Kong SAR, China
e-mail: kcshih@hku.hk

V. Vardhanabhuti
Department of Diagnostic Radiology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

E. Tsougenis
Artificial Intelligence Lab, Hospital Authority, Hong Kong SAR, China

Key Summary Points

The public specialty ophthalmic clinics in Hong Kong, under the Hospital Authority, receive tens of thousands of referrals each year. Triaging these referrals incurs a significant workload for practitioners and the other clinical duties.

Machine learning (ML) has been shown to improve the efficiency of various medical workflow, including triaging, by both reducing the workload and increasing accuracy in some cases.

This study proposes a general framework for developing, deploying and evaluating a machine learning-based triaging algorithm in a clinical setting.

We hope this framework, albeit not exhaustive, can act as a foundation to accelerate future pilot studies and deployments.

INTRODUCTION

Timely and accurate triaging to ophthalmology specialist outpatient clinics (SOPC) is vital in holistic patient care and an efficient public health system. Currently, outpatient referrals from primary care are sorted into one of three categories by healthcare professionals in Hong Kong: urgent, semi-urgent, and stable cases. Unfortunately, the size of workforce in Hong Kong SOPC is generally not adequate for the great demand from the public [33]. For instance, the median waiting time for stable cases at Kowloon Central Cluster is 64 weeks, and the longest waiting time is up to 148 weeks [17]. With such a huge workload, the triaging process can become very time-consuming, and workplace burnout and oversight can increase [18].

Furthermore, with limited resources for ophthalmology clinic appointments, the more accurate a triage system is, the more patients with potentially correctable sight-threatening pathology can be attended to for timely and effective intervention. A study in England reported that a median delay in care of 22 weeks resulted in permanently reduced visual acuity (VA) in 72% of patients and deterioration in the visual field (VF) in 23% of patients [13]. These show the urgent public health need for novel and efficient solutions to improve the waiting time and accuracy of triage systems. Teleophthalmology models of care have been deployed in England and Singapore and shown to enable more targeted referrals from primary care to specialists [4, 27]. However, infrastructure and human resources are still significant constraints.

For years researchers have been exploring the possibility of using artificial intelligence (AI) as a decision support system for medical triage [32]. The use of intelligent systems can potentially alleviate the aforementioned human resources constraints in processing referrals. With the advancement in AI and its related technologies, more and more departments worldwide are starting to evaluate and even adopt AI in their triaging workflow to increase the accuracy and decrease the workload of the clinical staff. For instance, the audiology department at Mayo Clinic in Florida, USA deployed a triage AI algorithm at their clinic. The algorithm was evaluated on three metrics: the algorithm's accuracy, savings to clinicians' time as compared to manual triaging, and the average number of appointments saved. They found an approximately 20% drop in dizziness referrals to otolaryngology and led the clinic to find significant over-referrals of dizziness to otolaryngology [10]. Other examples include dermatology, emergency departments and COVID-19 triage. In dermatology, a meta-analysis [46] found that the accuracy of computer-aided diagnosis for melanoma detection is comparable to that of experts at a sensitivity of 0.74 and specificity of 0.84, but noted uncertainties in real-world applicability owing to overfitting and uncorrected bias in some of the studies. In emergency departments, a systemic review found an improvement in the health professionals' decision-making, thereby leading to better clinical management and patient outcomes [25]. In COVID-19 triage, a hospital in

Switzerland deployed a machine learning model in predicting severe outcomes and achieved an AUC of 0.94 in both retrospective and prospective cohort studies [35].

Artificial intelligence is a vast field of study, and machine learning (ML) is a subset of AI which allows an algorithm to detect patterns within data without explicit instructions. Deep learning (DL) is a popular subset within ML algorithms. Its popularity in recent years is due to advancements in computation power and big data, making these algorithms computationally tractable even with a large amount of data. The DL architecture most commonly used in the medical imaging field includes convolutional neural network (CNN), which is partly inspired by the structure of our biological visual system. Another widely used architecture is the transformer [45], commonly used in natural language understanding (NLU)[1] for biomedical texts processing. A general understanding of all these architectures helps in planning for a framework as they confer different trade-offs, e.g. accuracy, computation power needed, explainability.[2]

Despite encouraging pilot studies, significant difficulty exists between model development and translation into clinical application and outcome [39]. This article aims to propose a production ML-based framework for triaging referrals to ophthalmic outpatient clinics in Hong Kong, with the end goal of reducing triaging workload and improving patients' wellbeing. This paper will go through the data collection process, architecture, deployment and operational aspects of such a system and discuss possible impacts and limitations. Lastly, we hope this paper can provide useful insights to bridge the gap between ML and clinical utility.

METHODOLOGY

The study involved searching relevant literature in artificial intelligence in the biomedical field. Google Scholar, PubMed and Microsoft Academic databases were explored in the search. Keywords included "ophthalmology", ["machine learning" OR "artificial intelligence" OR "deep learning"], ["triaging" OR "referrals"] and "teleophthalmology". Only English language journals were included. A review of the literature was performed to assess the architecture, deployment and evaluation metrics. The clinical impacts and limitations of the studies were noted. Afterwards, a conceptual framework for machine learning-aided triaging was developed based on the current system in public specialty ophthalmology clinics in Hong Kong. This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

FRAMEWORK AND DISCUSSION

An end-to-end clinical framework for deploying intelligent systems should include data collection and pre-processing protocols, data annotation, system architecture, deployment strategies and evaluation metrics. Our proposed framework uses colour fundus photos (CFPs) and referral synopses as the inputs since pilot studies have shown their utility in triaging [41] and grading important diseases, including diabetic retinopathy (DR) [14], age-related macular degeneration (AMD) [5] and urgent cases such as glaucoma [6]. Research in emergency department triaging has also found that textual data is significant for the accuracy of a triaging ML network [31]. Other imaging studies such as optical coherence tomography might be useful, but generally not performed in the primary sector, and therefore is not included. For pilot studies, a portion of the referral can be processed by the framework delineated in Figs. 1 and 2, and the results compared to the current protocol via metrics including accuracy, consultation time saved and waiting time reduced.

[1] NLU focuses on the meaning and context of the texts, while natural language processing (NLP) encompasses a much wider gamut of tasks including speech recognition, part-of-speech tagging, structure extraction etc.

[2] Explanability refers to the ability to explain or visualise how a network arrives at the output.

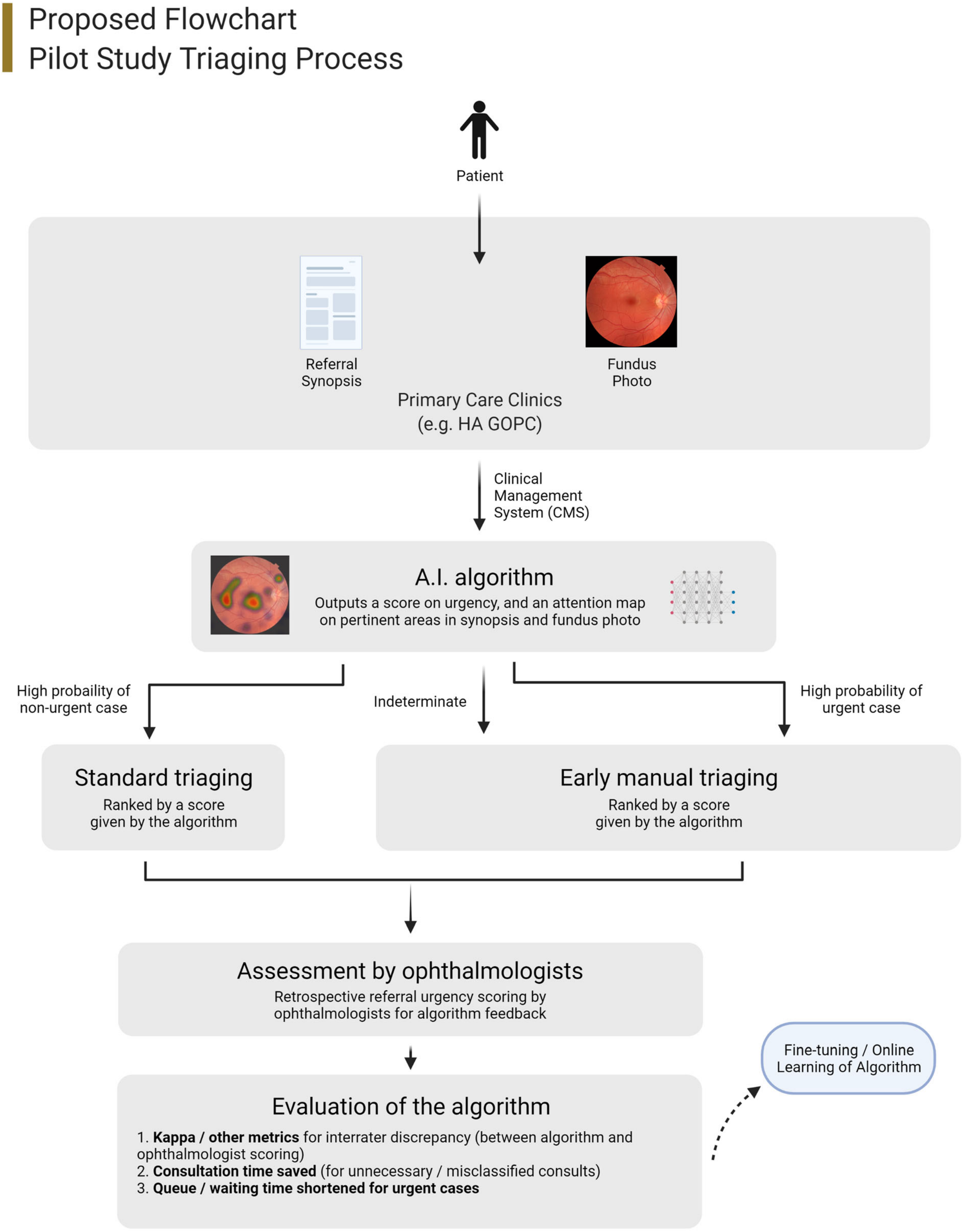
Proposed Flowchart
Pilot Study Triaging Process
Patient
Referral Synopsis
Fundus Photo
Primary Care Clinics (e.g. HA GOPC)
Clinical Management System (CMS)
A.I. algorithm
Outputs a score on urgency, and an attention map on pertinent areas in synopsis and fundus photo
High probaility of non-urgent case
Indeterminate
High probability of urgent case
Standard triaging
Ranked by a score given by the algorithm
Early manual triaging
Ranked by a score given by the algorithm
Assessment by ophthalmologists
Retrospective referral urgency scoring by ophthalmologists for algorithm feedback
Evaluation of the algorithm
1. Kappa / other metrics for interrater discrepancy (between algorithm and ophthalmologist scoring)
2. Consultation time saved (for unnecessary / misclassified consults)
3. Queue / waiting time shortened for urgent cases
Fine-tuning / Online Learning of Algorithm

Fig. 1 Overview of the proposed framework

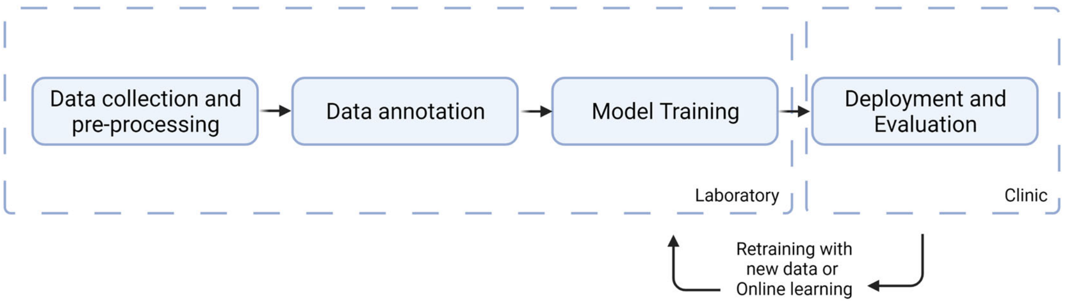
Data collection and pre-processing
Data annotation
Model Training
Deployment and Evaluation
Laboratory
Clinic
Retraining with new data or Online learning

Fig. 2 Development flow chart

Data Collection, Cohort Selection, and Data Pre-processing

The first step for developing a machine learning-based system is to devise a protocol for collecting data. In healthcare, the development team should coordinate and ensure compliance with the privacy law in the region—for Hong Kong, it is the Personal Data (Privacy) Ordinance. Data should undergo de-identification, and patients' permissions should be acquired. Furthermore, utilising CFPs from other sources such as APTOS [2], Messidor-2 [26], and Eye-PACS [12] datasets for pre-training can reduce training time and increase accuracy.

Data should undergo pre-processing. Pre-processing may encompass many steps, including data cleaning, standardisation, image or text processing, and defining protocols for data augmentation.

Data cleaning refers to removing poor quality data and thus reducing the possibility of garbage in, garbage out (GIGO).[3]

Image processing includes techniques such as noise removal filters, feature extraction filters and colour manipulation. They have been shown to visually emphasise important features such as optic disc boundaries and blood vessels [16]. However, whether these techniques improve network detection rate significantly is still unclear and will require further research. Textual data processing includes tokenisation and padding. Previous research [41] demonstrated good accuracy with sub-word tokenisation and word-sequence-independent methods (e.g. ANN, random forest), achieving an AUC of 0.83 and accuracy of 0.81 in categorising urgent vs non-urgent ophthalmology referrals. With these methods of pre-processing, the referral texts do not have a defined structure and therefore can contain any information. The research identified word stems like 'IOP', 'vision' and 'urgent' being more significant, and therefore it is likely this information is more useful in categorisation. *Huggingface* [48] is an open-source library for tokenisers and pre-trained models in natural language processing (NLP). For pilot studies, biomedical texts tokenisers and pre-trained models from *Huggingface* can be used for the foundation.

The data format should be standardised before annotation and training. Since medical data frequently contains missing data, imputations should also be performed.

Imbalance among different data sources (e.g. different clinics, equipment) should be identified by performing exploratory data analysis (EDA). Multiple factors, including patient demographics, labels, image quality and outliers, should be thoroughly explored. Furthermore, note if the distribution of labels is imbalanced across data sources. For instance, if two data sources have a significant difference in the ratio of positive to negative samples, the algorithm might learn subtle differences (e.g. lighting, artefacts) between the two data sources instead of learning the features of the CFPs for prediction. Moreover, positive and negative samples should be inspected to look for any feature leakage. For example, positive samples acquired directly from centres might have annotations made by optometrists or ophthalmologists on the CFPs.

[3] GIGO refers to when a network is fed "garbage" data, it will have poor predictive power.

Data augmentations are techniques that can increase the amount of data by adding modified copies of the original data (e.g. rotation in images, random deletion in texts) [36]. They can act as regularisation and an approach to unbalanced datasets.

Data Annotation

For passive annotation, the labels can be acquired from past triaging records. International Classification of Diseases 10th Revision (ICD-10) medical codes and interoperability with the clinical management system (CMS) can be leveraged for faster passive annotation. Since all cases are digitalised and documented with ICD-10 codes in Hong Kong Hospital Authority, we can trace and extract the final diagnosis and procedures done for each referral case. The optimal referral category (urgent, semi-urgent, stable) of the referrals can therefore be inferred by experts. For training data, we can retrospectively retrieve past referral cases and use the diagnoses and procedures for annotations. For evaluation, we can trace the final diagnosis and procedures done and derive the optimal referral category.

However, if we wish to annotate CFPs separately, we might require manual annotation. Since medical data annotation requires domain experts, we cannot use crowdsourcing services like scale.ai or Amazon Mechanical Turk. Researchers have proposed a crowdsourcing framework for medical datasets and can serve as a reference and guidelines in protecting patients' privacy [52].

Research has shown considerable grader variability by medical professionals in the analysis of CFPs and adjudication,[4] instead of taking a simple majority, is a better way to annotate unlabelled data [22]. This nevertheless incurs a heavier workload and might be hard to adopt. For medical AI development, a common bottleneck is the resource-intensive process of medical image and data annotation. Therefore, for confident samples, simple majority or individual expert can be used while reserving adjudication for ambiguous cases.

Model Architecture and Training

A model architecture is arrived at mainly by trial and error. However, we can try to provide a framework based on literature review. Before implementing the architecture and training, the dataset should be split into a development set and test set. For the development set, cross-validation, bootstrap, or hold-out validation strategies can be used depending on the number of samples and computation power of the project [21, 50]. For grading of diabetic retinopathy, previous studies have used both cross-validation and hold-out validation to achieve the state of the art (SOTA) results [43, 44].

Network architecture changes rapidly from time to time in the field of artificial intelligence. Networks used before in image classification tasks include small networks like VGG16 [20] to large networks like U-Net and Inception-v3 [44]. The current SOTA on image classification is EfficientNet [42], whose pre-trained weights on ImageNet [9] are publicly available. For NLU tasks, networks in the literature on biomedical texts span from convolutional neural networks [41] to transformers (e.g. BERT and its variants) [51].

Owing to the complexity of fundus photos and referral synopsis, transfer learning and self-supervised learning [40] can be considered before the main training routine. Transfer learning can reduce the training time and increase the final accuracy [47], while self-supervised learning is suitable for inadequate annotated data or a large amount of data.

In transfer learning, we "transfer" the weights of a previously trained network to our network and substitute the last few layers to our needs, followed by "fine-tuning" the network with our data. Previous research has shown that transfer learning greatly reduces time and improves accuracy [20] in tasks similar to ours. For instance, if we use EfficientNet, pre-trained weights on ImageNet can be used as the foundation, followed by pre-training on third-party databases (e.g. APTOS), and finally fine-tuned

[4] Adjudication refers to a group of experts deliberating together on sample annotation.

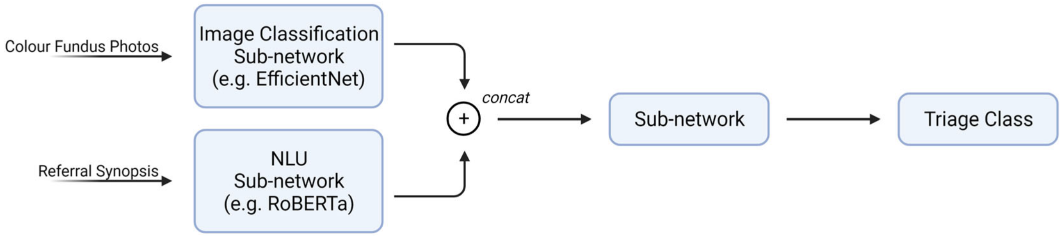
Colour Fundus Photos
Referral Synopsis
Image Classification Sub-network (e.g. EfficientNet)
NLU Sub-network (e.g. RoBERTa)
+
concat
Sub-network
Triage Class

Fig. 3 Proposed network architecture

on the data we acquired from our locality. Ensembling[5] can also be considered.

If we use both CFPs and referral synopsis as the input to our network, our architecture will need to combine these two inputs to generate the final prediction. We can first obtain the baseline by training two networks using only CFPs or referral texts respectively. To combine the two inputs, we can add a *concat* layer after the respective pre-trained network. Afterwards, we train the whole network. Past research has shown that such a technique can improve the performance of a network considerably [11, 37]. A possible network architecture is illustrated in Fig. 3.

Before training, the loss function and relevant metrics (e.g. AUC, weighted Cohen's kappa) should also be defined. Heavier weighting to specific groups (e.g. urgent cases) can be considered as they incur a significant public health risk. Since the triage levels (i.e. stable, semi-urgent, urgent) are ordinal, we can use a linear unit for the final output. Alternatively, we can have three logit outputs as the probability of the respective classes. If the threshold is not met on all three outputs, the network deems the case ungradable.

Model Deployment and Clinical Evaluation

An AI system may perform well in the laboratory, but faces significant challenges in a clinical environment or even fail to have any clinical impact when deployed [15]. Google's attempt at using AI to screen for diabetic retinopathy in Thai clinics [3] is an example of the immense challenges faced when an algorithm is deployed in the real world. For instance, blurry images that are human-readable were rejected by the system and caused frustration with nurses, and poor internet connections cause delays in screening at clinics. Therefore, a human-centred approach and evaluation are needed when deploying a medical AI system.

Logs should be kept and inspected for the deployed system, e.g. percentage of images or texts that are ungradable. A proper channel for doctors' and nurses' feedback should be established, and on-site surveys should be performed from time to time. It is noted in previous research that human and societal factors have as much impact as the accuracy of the algorithm on the clinical efficacy of the model. Therefore, a successful clinical AI decision tool requires a user-centric approach in post-deployment improvements.

To evaluate clinical impact, we can track metrics including clinician's time saved, the average number of appointments saved and time to referral. Metrics should reflect the clinical utility the tool contributed, instead of merely the model's performance. Nevertheless, the model's performance should be tracked as trained static models are known to degrade over time [1] because of concept shifts such as changes in imaging equipment, improvement in image qualities, or changes in the underlying distribution. Misclassification should also be noted and inspected.

For better integration, interoperability with the clinical management system/hospital information system (HIS) can be considered. The model can be hosted either on-site, if equipment allows, or on the cloud [19].

[5] Ensemble methods is a technique that combines several models to produce a better predictor e.g. simple average over multiple outputs.

Infrastructure should be planned with system experts, and various researchers have suggested blueprints [23].

Model Explainability

Model explainability is vital in medical AI as decisions should be explainable to both the practitioners and patients. Previous research has used attention mechanism [29] and Grad-CAM [7] to show the pertinent areas in images and texts. Other methods include SHAP (SHapley Additive exPlanations) [24] and Lime [30]. We can visualise which part of the text or image the network paid most attention to, to generate the output through these methods.

LIMITATIONS AND CHALLENGES

Several challenges can be foreseen. For practical challenges, interoperability with existing HIS and workflow is useful but challenging to establish. Clinical guidelines have to be adapted, which involves a lot of stakeholders and deliberation. Model explainability is still in its infancy for more complicated networks and might not satisfy practitioners' and patients' expectations. For technical challenges, previous works on CFPs mainly focus on a single disease like diabetic retinopathy or glaucoma, but a referral system will need to recognise multiple conditions. An acceptable model might take a long time to develop, and resources might be limited, impacting development and deployment. For legal challenges, legal liability is an area of concern and approaches to addressing them are multifaceted [34]. For pilot studies, a human should monitor the system. Incorporating human experts in AI systems is known as human-in-the-loop (HITL), and limitations of both humans-only and AI-only systems can be addressed. Limitations for AI-only systems include inaccuracies in rarer conditions, and limitations for human experts include fatigue [38]. HITL has been employed in chest radiograph diagnosis and was superior to AI-only or human expert-only systems, with HITL model achieving an AUC of 0.840, the experts-only system achieving 0.763 and AI-only system achieving 0.685 [28]. Also, further collaboration with legal experts should be considered if a system is to be deployed. While HITL systems still require human supervision and the efficiency increment varies from case to case, HITL systems have been shown to improve both accuracy and efficiency in radiology reporting [49] and general medical triaging [8]. The degree of human involvement can also be varied according to the confidence of the network, further improving efficiency.

CONCLUSION

This framework for a machine learning-aided triaging system in ophthalmology clinics in Hong Kong has been developed after reviewing the local situation and literature. This article provided a simple overview of the possible strategies for data collection, pre-processing, data annotation, system architecture, deployment strategies, evaluation metrics, limitations and challenges of such a system. In deploying and utilising medical AI systems, we should pay equal attention to both the accuracy of the network as well as the actual clinical utility when deployed. Although future work is needed to validate the proposal listed in this article, we hope this framework will be helpful as a foundation for development, pilot testing and deployment for referral system in both ophthalmic clinics and other specialty clinics.

ACKNOWLEDGEMENTS

Figures created with BioRender.com.

Funding. No funding or sponsorship was received for this study or publication of this article. The journal's Rapid Service Fee was funded by the authors.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Authorship Contributions. Yalsin Yik Sum Li and Kendrick Co Shih were involved in study design, data collection, data analysis, manuscript writing and editing. Varut Vardhanabhuti, Efstratios Tsougenis and Wai Ching Lam were involved in data collection, data analysis, manuscript writing and editing.

Disclosures. The following authors confirm that they have nothing to disclose: Yalsin Yik Sum Li, Varut Vardhanabhuti, Efstratios Tsougenis, Wai Ching Lam, Kendrick Co Shih.

Compliance with Ethics Guidelines. This article is based on previously conducted studies and does not contain any new studies with human participants or animals performed by any of the authors.

Data Availability. Data sharing is not applicable to this article as no datasets were generated or analysed during the current study. The authors agree to make all materials, data and associated protocols promptly available to readers without undue qualifications in material transfer agreements.

Declaration of Interest. The authors alone are responsible for the content and writing of the paper. This manuscript has not yet been published and is not being simultaneously considered elsewhere for publication.

REFERENCES

1. Adam GA, Chang CHK, Haibe-Kains B, Goldenberg A. Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. Proceedings of the 5th Machine Learning for Healthcare Conference. PMLR 2020;126:710–31.

2. APTOS. APTOS 2019 blindness detection. Retrieved from Kaggle. 2019. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed 4 July 2021.

3. Beede E, Baylor E. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: Conference on human factors in computing systems. Retrieved from Healthcare AI systems that put people at the center. 2020, 4 25. pp. 1–12. https://www.blog.google/technology/health/healthcare-ai-systems-put-people-center/. Accessed 4 July 2021.

4. Borooah S, Grant B. Using electronic referral with digital imaging between primary and secondary ophthalmic services: a long term prospective analysis of regional service redesign. Eye (Lond). 2013;27(3):392–297.

5. Burlina PM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol. 2017;135(11):1170–6.

6. Chen XX, Xu Y, Wong DWK, Wong TY, Liu J. Glaucoma detection based on deep convolutional neural network. Annu Int Conf IEEE Eng Med Biol Soc. 2015;2015:715–18.

7. Chetoui M, Akhloufi MA. Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets. J Med Imaging. 2020. https://doi.org/10.1117/1.JMI.7.4.044503.

8. Delshad S, Dontaraju VS, Chengat V. Artificial intelligence-based application provides accurate medical triage advice when compared to consensus decisions of healthcare providers. Cureus. 2021;13(8):e16956.

9. Deng J, Dong Q, Socher R, Li L, Li K, Li FF, ImageNet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–55.

10. McCaslin, D. 20Q: Using artificial intelligence to triage and manage patients with dizziness—The Mayo Clinic experience. AudiologyOnline, Article 26880. Retrieved from www.audiologyonline.com. (2020). Accessed 4 July 2021.

11. Du P, Li X and Gao Y. Employ Multimodal Machine Learning for Content Quality Analysis, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, pp. 2658–61.

12. EyePACS. EyePACS. Retrieved from EyePACS. 2021. http://www.eyepacs.com/. Accessed 4 July 2021.

13. Foot B, MacEwen C. Surveillance of sight loss due to delay in ophthalmic treatment or review: frequency, cause and outcome. Eye (Lond). 2017;3(5): 771–5.

14. Gulshan VP. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10.

15. Hartswood M, Procter R. 'Repairing' the machine: a case study of the evaluation of computer-aided detection tools in breast screening. ECSCW. Berlin: Springer; 2003. p. 375–94.

16. Hashim FA, Salem NM and Seddik AF. Preprocessing of color retinal fundus images, 2013 Second International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), 2013, pp. 190–3

17. Hospital Authority. Waiting Time for New Case Booking atEye Specialist Out-patient Clinics. Hospital Authority Public Website. https://www.ha.org.hk/haho/ho/sopc/dw_wait_ls_eng.pdf. Accessed 30 May 2021.

18. Jin Wen YC. Workload, burnout, and medical mistakes among physicians in China: a cross-sectional study. BioSci Trends. 2016;10(1):27–33.

19. Kern C, Fu DJ. Implementation of a cloud-based referral platform in ophthalmology: making telemedicine services a reality in eye care. Br J Ophthalmol. 2020;104:312–7.

20. Khalifa N, Loey M. Deep transfer learning models for medical diabetic retinopathy detection. Acta Inform Med. 2019;27(5):327–32.

21. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. 1995; 2(12): 1137–43.

22. Krause J, Glushan V. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology. 2018;125(8):1264–72.

23. Leiner T, Bennink E. Bringing AI to the clinic: blueprint. Insights Imaging. 2021;12(1):1–11.

24. Lundberg SM, Lee S-I. A unified approach to interpreting model. Long Beach: NIPS; 2017.

25. Marta Fernandes SM. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. Artif Intell Med. 2020;102: 101762.

26. MESSIDOR-2 DR Grades. Retrieved from Kaggle. 2018, 7 3. https://www.kaggle.com/google-brain/messidor2-dr-grades. Accessed 4 July 2021.

27. Nguyen HV, Gavim SWT. Cost-effectiveness of a national telemedicine diabetic retinopathy screening program in Singapore. Ophthalmology. 2016;123(12):2571–80.

28. Patel BN, Rosenberg L, Willcox G. et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. npj Digit. Med. 2019;2:111.

29. Poplin R, Varadarajan AV. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3): 158–64.

30. Ribeiro MT, Singh S. "Why should I trust you?": explaining the predictions of any classifier. San Francisco: KDD; 2016.

31. Roquette BP, Nagano H. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. Neural Netw. 2020;126:170–7.

32. Sadeghi S, Barzi A, Zarrin-Khameh N. Decision support system for medical triage. Stud Health Technol Inform. 2001;81:440–442.

33. Schoeb V. Healthcare service in Hong Kong and its challenges. China Perspect. 2016;2016(4):51–8.

34. Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. Int J Law Inf Technol. 2019;27(2):171–203.

35. Schöning V, Liakoni E, Baumgartner C, et al. Development and validation of a prognostic COVID-19 severity assessment (COSA) score and machine learning models for patient triage at a tertiary hospital. J Transl Med. 2021;19(1):56.

36. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48.

37. Soguero-Ruiz C, Hindberg K, Mora-Jiménez I. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. J Biomed Inform. 2016;61:87–96.

38. Stec N, Arje D, Moody AR, Krupinski EA, Tyrrell PN. A systematic review of fatigue in radiology: Is It a Problem?. AJR Am J Roentgenol. 2018;210(4): 799–806.

39. Strickland E. IBM Watson: heal thyself: How IBM overpromised and underdelivered on AI health. IEEE Spectr. 2019;56(4):24–31.

40. Taleb A, Loetzsch W, Danz N, Severin J, Gaertner T, Bergner B, Lippert C. 3D self-supervised methods for medical imaging. Vancouver: NeurIPS; 2020.

41. Tan YB. Triaging ophthalmology outpatient referrals with machine learning: a pilot study. Clin Exp Ophthalmol. 2020;48(2):169–73.

42. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. International conference on machine learning. ArXiv. 2019. pp. 6105–14.

43. Tymchenko B, Marchenko P, Spodarets D. Deep learning approach to diabetic retinopathy detection. arXiv:2003.02261. 2020.

44. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm. JAMA. 2016;316(22):2402–10.

45. Vaswani A, Shazeer N. Attention is all you need. Long Beach: NIPS 2017; 2017.

46. Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of computer-aided diagnosis of melanoma. JAMA Dermatol. 2019;155(11):1219.

47. Wan S, Liang Y. Deep convolutional neural networks for diabetic retinopathy detection by image classification. Comput Elect Eng. 2018;72:274–82.

48. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771. 2019.

49. Wu JT, Syed A, Ahmad H, et al. AI accelerated human-in-the-loop structuring of radiology reports. AMIA Annu Symp Proc. 2021;2020:1305–1314.

50. Yadav S, Shukla S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In: 2016 IEEE 6th international conference on advanced computing (IACC). 2016. pp. 78–83.

51. Yang X, Bian J. Clinical concept extraction using transformers. J Am Med Inf Assoc. 2020;27(12): 1935–42.

52. Ye C, Coco J. A crowdsourcing framework for medical data sets. AMIA Jt Summits Transl Sci Proc. 2018;2018:273–80.