Triaging ophthalmology outpatient referrals with machine learning: A pilot study

Authors: Ian Tan[1], Stephen Bacchi[1], Robert Casson[1], Dinesh Selva[1], Weng Onn Chan[1]

Affiliations:

[1]Royal Adelaide Hospital**,** Adelaide, Australia

Correspondence:

Ian Tan

Royal Adelaide Hospital

Ophthalmology Department

Port Road, Adelaide, 5000

Tel: 70740000

Email: ian.tan1100@gmail.com

Cover title: Deep learning triaging for ophthalmology referrals

KEY WORDS: Deep Learning, Natural Language processing, machine learning, ophthalmology

ABSTRACT:

Importance: Triaging of outpatient referrals to ophthalmology services is required for the maintenance of patient care and appropriate resource allocation. Machine learning (ML), in particular natural language processing, may be able to assist with the triaging process.

Background: The aim of this study is to determine whether ML can accurately predict triage category based on ophthalmology outpatient referrals.

Design: Retrospective cohort study

Participants: The data of 208 participants was included in the project. This data included 118 category one referrals, 61 category two referrals and 29 category three referrals.

Methods: The synopses of consecutive ophthalmology outpatient referrals at a tertiary hospital were extracted along with their triage categorisations. Following pre-processing, ML models were applied to determine how accurately they could predict the likely triage categorisation allocated. Data was split into training and testing sets (75%/25% split). ML models were tested on an unseen test set, after development on the training dataset.

Main Outcome Measure: Area under the receiver operator curve (AUC) for category one vs. non-category one classification.

Results: For the main outcome measure, convolutional neural network (CNN) provided the best AUC (0.83) and accuracy on the test set (0.81), with the artificial neural network (AUC 0.81 and accuracy 0.77) being the next best performing model. When the CNN was applied to the classification task of identifying which referrals should be allocated a category one vs. category two vs. category three priority, a lower accuracy was achieved (0.65).

Conclusion and Relevance: These results demonstrate that ML may be able to accurately assist with the triaging of ophthalmology referrals. Future studies with data from multiple centres and larger sample sizes may be beneficial.

MANUSCRIPT:

1. Introduction

Correct triaging to ophthalmic services is integral to patient care and appropriate resource allocation. To ensure that patients with urgent issues are seen in a timely fashion, outpatient referrals are usually manually sorted into one of several categories, from most urgent to least urgent. In many centers this process is currently performed by both nurse and medical practitioners. Triaging of referrals is integral to patient care because appointments for ophthalmology review are limited, and patients with potentially correctable sight-threatening pathology may need to be seen within a given window of time to enable effective intervention. The current process involved in the categorization of referrals may be time consuming and there is the potential for error when the person triaging is less experienced.

Machine learning (ML) may be considered the use of computer programs to detect patterns within data and perform tasks, without having been explicitly instructed how to do so (1). Deep learning (DL) is a subset of ML that focusses largely on the use of artificial neural networks (ANN), and associated architectures such as convolutional neural networks (CNN). While most commonly known for their use in image analysis (2), CNN can be applied to language. The application of ML to human language may be described as natural language processing (NLP), which is being researched in various clinical disciplines (3).

There is a great deal of interest in potential applications of DL to ophthalmology (4). Such DL ophthalmology applications have been reviewed previously (5). The majority of these studies have focused on image interpretation (such as fundus photographs and visual field analysis). We could not identify any studies that specifically applied DL or NLP to the issue of ophthalmology referral triage.

The aim of this project was to use retrospectively collected outpatient ophthalmology referrals to determine how effectively DL NLP can (Aim 1) identify referrals requiring a “category one” (urgent) prioritization, and (Aim 2) emulate human triaging across all three categories.

2. Method

Data collection

Data were collected from consecutive referrals to the Royal Adelaide Hospital Ophthalmology Outpatient Department during the period: give dates. Referrals for these patients had all been made within the previous 24 months. A synopsis of the referral, triage categorisation, and the source of the referral was recorded.

Pre-processing

Individuals for whom there was incomplete referral data or outcome data were excluded. If text was to be classified by a word-sequence-independent method (such as an ANN or Random Forest), negation detection was applied. Referral text punctuation was removed. Word stemming was performed, followed by tokenisation. The least frequently appearing 1% of words were excluded from the corpus.

Prior to analysis by a CNN, token sequences were padded to provide a consistent sequence length. Prior to analysis with word-sequence-independent methods, count vectorisation was performed.

Data was then randomly split into training and testing sets (75%/25%). This split was performed once.

Classifier development

Using the training set, models were trialled using 5-fold cross-validation. Variable neural network architectures were trialled on the training data. Initially, basic architectures with few nodes and hidden layers were used. With subsequent iterations, further layers were added until an optimal accuracy was achieved on the training data. Hyperparameter tuning was conducted on the training data.

The final CNN architecture employed was: an embedding layer, dropout layer, convolutional layer, maximum pooling layer, and then 5 dense hidden layers (nodes varying from 512 to 128).

Model assessment

The developed models were then used to predict the categorisations of the hold-out test set. In binary classification tasks (Aim 1) Youden's index was used to select the cut-off score for each model. Initially, all models were used to predict the binary outcome in Aim 1 (category one vs non-category one). In Aim 1 the primary outcome was area under the receiver operator curve (AUC). Other outcomes assessed included accuracy, F1 score, positive predictive value (PPV), negative predictive value (NPV), sensitivity and specificity. Examples of results using different cut-off scores, demonstrating high sensitivity or high specificity, were generated for the best performing model.

The best performing model on Aim 1 was then employed to predict the actual triage category (category one vs two vs three) assigned to each referral (Aim 2). In Aim 2 the primary outcome was classification accuracy.

Due to the pilot nature of the study, no statistical tests were conducted to demonstrate superiority of one model as compared to another.

Institutional review

This project was submitted to the relevant institutional review board and considered exempt from approval (R20190108).

3. Results

Participant and referral characteristics

Data from 208 participants were included in this study. There were 118 category one referrals, 61 category two referrals and 29 category three referrals. These categorisations were allocated by a senior nurse practitioner (with more than 15 years of clinical experience) whose job it is to triage such referrals. Ninety-three of the participants were male (44.7%), and the average age of the participants was 57.7 (SD 18.6) years. SD is always just a positive number

The mean length of referral synopsis was 68.1 words (IQR 25-93, range 2-293 words). Referral sources included general practitioners (51, 24.5%), optometrists (57, 27.4%), specialists (98, 47.1%), and the emergency department (2, 1.0%). The referrals included both internal referrals, from within the tertiary hospital (64, 30.8%), and external referrals (144, 69.2%).

Identification of referrals requiring category one prioritisation

The CNN provided the best AUC (0.83) and accuracy on the test set (0.81) (see Table 1 and Figure 1). The next best performing models were the ANN (AUC 0.81 and accuracy 0.77) and logistic regression models (AUC 0.79 and accuracy 0.77). The Random Forest (AUC 0.77 and accuracy 0.73) and Decision Tree (AUC 0.58 and accuracy 0.6) models achieved lower accuracies.

When different cut-off scores were employed for the CNN model, high specificities or high sensitivities were able to be achieved, at the expense of overall accuracy (see Table 1).

Coefficients of the most strongly predictive words were extracted from the Logistic Regression model to gauge the words on which the models may be placing the most emphasis. The word stems that were most predictive of category one were "urgent", "vision", "IOP", "disc" and "left". The word stems that were most predictive of non-category one classification were "cataract", "diabet", "le", "mr", and "diseas".

Emulation of human triaging

When the CNN was applied to the multi-task classification task of identifying which referrals should be allocated a category one vs category two vs category three priority, a significantly lower accuracy was achieved (0.65).

4. Discussion

Our results demonstrate that ML, in particular DL, can accurately assist with the triaging of ophthalmology referrals. For example, it would seem feasible that a system could be developed that would flag certain referrals definitely requiring a category one prioritisation (high specificity/PPV). It is important to note that our current CNN model is achieved entirely based on text entry alone. Conceivably, a CNN model that accounts for multi-modal input such as patient demographics, source of referral, clinical images would achieve an even better prediction.

Lower accuracies were achieved when a multi-class classification task was attempted when trying to emulate human triaging (Aim 2). The reason for this result is likely due to small sample size. In the entire dataset, there were only 29 referrals to which a category three prioritisation was allocated. Higher accuracies could likely be achieved with larger sample sizes. The inclusion of words such as “left” in those with high predictive value likely represents a degree of overfitting. This is most likely to occur in studies with small sample sizes (6), and a larger sample size would likely help to correct this issue.

ML has previously been successfully applied to the task of triage in other fields. For example, ML has been shown to be accurate in the triage of COPD exacerbations based upon pre-defined categorical and continuous variables (7), as opposed to the clinical text used in this project. Our pilot study is distinct as we demonstrated with a relatively small sample size that NLP can accurately identify urgent referrals from the full unfiltered spectrum of clinical ophthalmology referrals instead of just a specific disease process. It should be noted that the proposed DL model would not be looking to replace verbal communication in urgent cases. However, it is possible that DL models may be able to effectively emulate, and therefore streamline and/or cross-check current triage processes. Once models have been trained, they could be implemented on a regular computer without excessive processing power requirements.

Due to the pilot nature of this study, the greatest limitation was low sample size. As discussed above, larger sample sizes would likely enable the development of significantly more accurate models. It should be noted that this study was conducted at a single centre and exclusively in English. The gold-standard for correct triaging in this study were the classifications allocated by nurse practitioners. An ideal gold-standard would involve double-marking with two individuals at a consultant level of training.

Future research in this area should endeavour to use larger sample sizes, consultant-level triage allocation, and data from multiple centres. The triaging of referrals to other specialty outpatient clinics, outside of ophthalmology, may also be investigated.

In summary, we have successfully demonstrated the utility of natural language processing in triaging ophthalmology referrals. Our CNN model achieved an AUC of 0.83

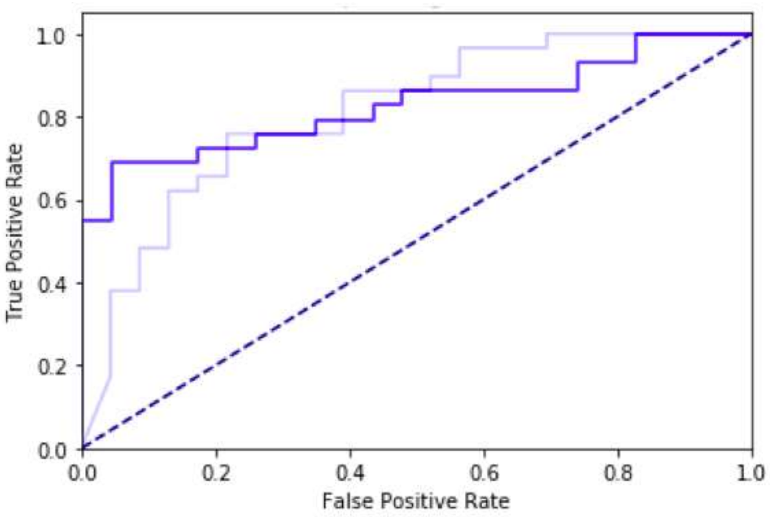
and accuracy of 0.81 in categorising urgent vs non-urgent ophthalmology referrals. Further prospective and comparative studies will be required to validate the model.

Table 1: Table demonstrating the results of machine learning applied to the identification of ophthalmology referrals that should be allocated a category one priority.

Model	Cut-off	AUC	TP	FN	TN	FP	Sensitivity	Specificity	PPV	NPV	F1 Score	Accuracy
CNN	Youden's index	0.83	20	9	22	1	0.69	0.96	0.95	0.71	0.80	0.81
CNN	High specificity	0.83	16	13	23	0	0.55	1.00	1.00	0.64	0.71	0.75
CNN	High sensitivity	0.83	28	1	4	19	0.97	0.17	0.60	0.80	0.74	0.62
ANN	Youden's index	0.81	22	7	18	5	0.76	0.78	0.81	0.72	0.79	0.77
Logistic Regression	Youden's index	0.79	21	8	19	4	0.72	0.83	0.84	0.70	0.78	0.77
Random Forest	Youden's index	0.77	20	9	18	5	0.69	0.78	0.80	0.67	0.74	0.73
Decision Tree	Youden's index	0.58	21	8	10	13	0.72	0.43	0.62	0.56	0.67	0.60

Abbreviations: CNN, convolutional neural network; ANN, artificial neural networks; AUC, area under the receiver operator curve; TP, true positive; FN, false negative; TN, true negative; FP, false positive; PPV, positive predicted value; NPV, negative predicted value.

Figure 1: ROC of CNN (dark blue – AUC 0.83) and ANN (light blue – AUC 0.81) in the prediction of ophthalmology referrals that should be allocated a category one priority.



5. References

1. Kapoor R, Walters SP, Al-Aswad LA. The current state of artificial intelligence in ophthalmology. Surv Ophthalmol. 2019;64(2):233-40.
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama. 2016;316(22).
3. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform. 2019;7(2):e12239.
4. Lee A, Taylor P, Kalpathy-Cramer J, Tufail A. Machine Learning Has Arrived! Ophthalmology. 2017;124(12):1726-8.
5. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2018.
6. Deo R. Machine Learning in Medicine. Circulation. 2015;123(1920-30).
7. Swaminathan S, Qirko K, Smith T, Corcoran E, Wysham NG, Bazaz G, et al. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. PLoS One. 2017;12(11):e0188532.