|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-I-041 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 7-8 May 2020 |
| **DOCUMENT** |
| **Source:** | Tencent Healthcare (China) |
| **Title:** | TG-Ophthalmo: Evaluation method and index of artificial intelligence glaucoma assisted screening system based on fundus image |
| **Purpose:** | Proposal |
| **Contact:** | Jianrong WuTencent Healthcare (Shenzhen)China | Email: edwinjrwu@tencent.com  |
| **Contact:** | Yanchun ZhuTencent Healthcare (Shenzhen)China | Email:  noonezhu@tencent.com  |
| **Contact:** | Man Tat Alexander NgTencent Healthcare (Shenzhen)China | Email:  alexanderng@tencent.com  |
| **Contact:** | Yajun ZhangTencent Technology (Shenzhen)China | Email:  yajunzhang@tencent.com  |

|  |  |
| --- | --- |
| **Abstract:** | The proposal in the context of TG-Ophthalmo is to provide a brief introduction of the definition, importance, procedure, requirements of data set construction and annotation in artificial intelligence assisted screening system based on fundus image. |

Algorithm and evaluation index

The AI glaucoma-assisted screening system based on fundus images should be evaluated based on the test set. The evaluation content includes false negatives and false positives, repeatability and reproducibility, robustness, sensitivity, specificity, and area under the operating characteristic curve（AUC）.

$$Sensitivity=\frac{Positive}{Positive+False negative}$$

$$Specificity=\frac{Negative}{Negative+False positive}$$

True positive is the data that is actually positive and correctly judged as positive; true negative is the data that is actually negative and correctly judged as negative; false positive is the data that is actually negative and wrongly judged as positive; Data that is negative. The sensitivity index measures the proportion of samples that are actually diseased and judged to be diseased; the specificity index measures the proportion of samples that are actually disease-free and judged to be disease-free.

ROC (Receiver-Operator Characteristic) curve, that is, the working characteristic curve or receiver operation characteristic curve: the probability value of using the AI model to predict the performance of the image glaucoma-like fundus, between 0-1, at 0.01 intervals to get different thresholds, namely Operation point, use this threshold to classify whether the image in the test set is glaucoma, and then obtain the sensitivity and specificity corresponding to the operation point, and plot the specificity as the horizontal axis and the sensitivity as the vertical axis to obtain ROC curve. The area between the ROC curve and the horizontal axis 0-1 is AUC.

In the calculation of sensitivity, specificity and AUC, the standard test set is randomly selected according to 80% of the patients' data. For example, 5,000 test sets are selected for 4000 cases, and each case is randomly selected with a picture for calculation, not less than Random sampling of 5 times was used for calculation, and the confidence intervals of the three indicators were counted.

Model repeatability: For sample cases with multiple fundus photos in the standard test set, calculate the output of the AI model for each case, repeat at least three sets, and calculate the Kappa coefficient between the two sets of results of the three sets. The resulting Kappa The average value of the coefficient is the repeatability of the model. The calculation of Kappa coefficient is based on confusion matrix, the specific calculation formula is as follows:

$$κ=\frac{p\_{o}-p\_{e}}{1-p\_{e}}$$

$$p\_{o}= \frac{\sum\_{i=0}^{c}M\_{i, i}}{n}$$

$$p\_{e}= \frac{\sum\_{i=0}^{c}\left(\sum\_{j=0}^{c}M\_{i,j} × \sum\_{j=0}^{c}M\_{j, i}\right)}{n^{2}}$$

Among them, c represents the prediction task of categories, M represents the confusion matrix between the two sets of predictions, which is a matrix of size c × c, and n represents the number of samples in each set of predictions. In the formula, $p\_{o} $is the observed agreement rate, and $p\_{e}$ is called the expected agreement rate, that is, the agreement rate of the two prediction results due to chance. The theoretical range of Kappa coefficient is between -1 and 1, but usually between 0 and 1. The Kappa value of 0.0-0.2 means there is almost no consistency, 0.21-0.39 means very low consistency, 0.4-0.59 means weak consistency, 0.6-0.79 means medium consistency, 0.8-0.9 means strong consistency, 0.9-1.0 means consistency is close to perfect.

Robustness: For the sample case containing a single fundus photo in the standard test set, or for the sample case containing multiple fundus photos, randomly select and fix the selected photo, and record the prediction result of the original image; then perform 5 on the original image Random left and right flips more than two times, random rotation within <= 10°, random cropping or translation within 5% of the image size, and record the image prediction results, calculating the Kappa coefficient between each group of prediction results and the original image prediction results The average value of the obtained Kappa coefficient reflects the robustness of the model against the corresponding image changes.

The specific requirements of the test algorithm model indicators: According to the statistics of diagnostic rate of fundus image of a large number of clinical ophthalmologists, considering all factors, we recommend that the verification results of the AI-assisted glaucoma screening system require specificity up to 85% and sensitivity up to 90% can be put into clinical use.

CDR evaluation

The optic disc diameter ratio (CDR) is between 0 to 1, and the CDR evaluation indicators include the root-mean-square error (i.e., standard deviation value) and R-squared value (coefficient of determination) between the predicted value and the calibration value.

The root mean square error can measure the deviation between the predicted value and the true value, and can reflect the accuracy of the measurement. The closer the rms error is to 0, the better the model is at predicting CDR values.

$$RMS=\sqrt{\frac{1}{Samples}\sum\_{}^{}(Predicted-Calibration )^{2}}$$

The R-squared value, also called the determination coefficient, is a statistical coefficient of the degree of fit between the regression prediction value and the calibration value. The R-squared value is between 0 to 1. The closer to 0, the closer the prediction result of the model is to random; the closer to 1, the better the fitting effect of the model regression prediction CDR value.

$$R^{2}=1-\frac{\sum\_{}^{}(Predicted-Calibration )^{2}}{\sum\_{}^{}(CMean-Calibration )^{2}}$$

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_