DC-CHECK (WORKED EXAMPLE 2)

Project Description: Deep Learning System for the Detection of Diabetic Retinopathy

Link:
[1] https://jamanetwork.com/journals/jama/fullarticle/2588763
[2] https://dl.acm.org/doi/pdf/10.1145/3313831.3376718

Citation:
[1] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, *316*(22), pp.2402-2410.
[2] Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P. and Vardoulakis, L.M., 2020, April. 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* (pp. 1-12).

DATA

Q1: How did you select, collect or curate your dataset?

- *Have you conducted forensics on the dataset (i.e. provenance)?*
- *Did you assess the pertinence of the dataset for the task?*
- *Is your dataset curation a once-off?*

Data curation

Macula-centered retinal fundus images were retrospectively obtained from EyePACS:

- United States and
- 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya).

Additional validation data collected

- EyePACS screening sites between May 2015 and October 2015. This data set did not overlap with the EyePACS data used in development.
- Publicly available Messidor-2 data set, obtained between January 2005 and December 2010 at 3 hospitals in France using a Topcon TRC NW6 nonmydriatic camera and 45° fields of view centered on the fovea.

Multiple camera types were used: including Centervue DRS, Optovue iCam, Canon CR1/DGi/CR2, and Topcon NW using 45° fields of view.

Q2: What data cleaning and/or pre-processing, if any, has been performed?

- *Do you need to "clean" your data?*
- *Do you have any missing data?*

- Input images were scale normalized by detecting the circular mask of the fundus image and resizing the diameter of the fundus to be 299 pixels wide.

- Images for which the circular mask could not be detected were excluded. This corresponded to 117 out of 128,175 on the development set, 17 out of 9,963 in EyePACS-1, and none in Messidor-2.

Q3: Has data quality been assessed?

- Completeness, diversity, label quality
- Easy, Hard examples etc

54 graders were used to validate and grade the images. All graders were US-licensed ophthalmologists or ophthalmology trainees in their last year of residency (postgraduate year 4).

A majority vote was taken if the image was considered referable I.e. of good quality, where the threshold is >0.5 agreement across raters.

Q4: Have you considered synthetic data?

Synthetic data has not been considered as:

- Data privacy is not an issue as de-identified data is available
- Data access is permitted. Hence, synthetic data is not required.
- Datasets are sufficiently large, hence no need to augment with synthetic data

TRAINING

Q5: Have you conducted a model architecture and hyperparameter search?

- *Have you specifically adapted the model for the task?*
- *Is there a potential to incorporate inductive biases?*

Model architecture:

Ensemble of neural networks used —> Inception v3 selected, pre-trained on ImageNet (adds inductive bias for better convergence)

Hyperparameters:

- Not specifically specified, besides the architecture

Q6: Does the training data match the anticipated use?

- *Can we anticipate the data the model will be applied on?*
- *Could the model be exposed to data from different distributions/domains?*

Data matching anticipated use:

- Yes, the models will be used in the same domain for diabetic retinopathy

Distribution/domain shift:

- There is likely a distribution shift if used in different countries or on different quality machines. e.g. Thailand - hence external validation needed

Q7: Are there different data subsets or subgroups of interest?

- *Are the subgroups identifiable?*
- *Have you assessed fairness/bias introduced by training?*

Subgroups/subsets of data:

- There are different subgroups of diabetic retinopathy: No diabetic retinopathy (45.4%), Mild diabetic retinopathy (25.9%), Moderate diabetic retinopathy (15.1%), Severe diabetic retinopathy (4.5%)

Training based on subgroup/subset:

- Currently, the models are optimized on average rather than per subgroup.

Q8: Is the data noisy, either in features or labels?

Presence of data noise:

- Noisy labels could be considered as variation in grading

Accounting for data noise:

- The quality control process accounts for this by filtering data points - hence no need to account for it during training

TESTING

Q9: How has the dataset been split for model training and validation?

- *Are you using a benchmark dataset?*
- *Is the splitting random?*

Data split

- Random 80-20 train-tuning split of the dataset

Q10: How has the model been evaluated (e.g. metrics & stress tests)?

- *What metrics are used to assess the model?*
- *Has the model been evaluated beyond average?*
- *Is there a potential to test specific aspects/sub-groups of the model (i.e. stress tests)?*

Metrics:

- Main: area under the receiver operating curve (AUC) generated by plotting sensitivity vs 1- specificity
- Secondary: Sensitivity as a high sensitivity is a prerequisite in a potential screening tool.

Assessment beyond average and/or subgroups:

- Additional sensitivity analyses were conducted for several subgroups:
  (1) detecting moderate or worse diabetic retinopathy only;
  (2) detecting severe or worse diabetic retinopathy only;
  (3) detecting referable diabetic macular edema only;
  (4) image quality; and
  (5) referable diabetic retinopathy on 2 data sets, each restricted to mydriatic and non-mydriatic images, respectively

DEPLOYMENT
Q11: Are you monitoring your model? - *Have you considered dimensionality in the monitoring?* - *Is there lag in ground truth feedback?*
A prospective study was carried out in Thailand & model performance was monitored based on this trial. No automated monitoring pipeline.
Q13: Do you have mechanisms in place to address data shifts? - *Have model updates been considered & when to retrain?* - *Does your system provide actionable feedback upon failure?* - *Are you able to characterize the type of shift?* - *Do system failures inform dataset updates?*
Model updates - Manual or automated model re-training pipelines have not been included **Failure feedback:** - The system provides feedback on failures, when the image is of low-quality **Characterizing the type of shift:** - We do not consider characterizing the type of shift **Failure informed dataset updates:** - The dataset is updated weekly, which while not failure informed would help to rectify issues, representing the latest data
Q14: Have you incorporated tools to engender model trust? - *Do you require predictive uncertainty estimates?* - *Does your ML system have explainability?* - *Does your system account for OOD inputs?* - *Do you assess issues of bias and fairness in deployment?*
Uncertainty estimates: - Uncertainty estimates of the predictions have not been included. **Explainability:** - No explicit model explainability **OOD:** - Yes, there is a gradeability detection system. - For patient safety reasons, it only assesses the highest-quality images. If an image has a bit of blur or a dark area, for instance, the system will reject it, even if it could make a strong prediction. **Bias/Fairness:** - Assessment of model bias or fairness gaps have not been provided