**Peer review**

Goteti, Shruti. 2024. "Automated Classification and Segmentation of Intracranial Hemorrhages Using 2D Convolutional Neural Networks and U-Net Architecture on Computed Tomography Scans." *Journal of High School Science* 8 (4): 381–413.

1.Was there a reason these datasets https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/data (2019 RSNA) or http://headctstudy.qure.ai/dataset (CQ500), or https://physionet.org/content/ct-ich/1.3.1/ (PhysioNet-ICH) were not utilized ? Some of them are significantly larger and appear in a number of literature CNN models (see this as an example: http://dx.doi.org/10.1136/).

2.I recommend using at least one of the above datasets as testing data. This will increase the robustness of your model. If you cannot use the data from any of these models for testing; then I recommend that you use your train:validate: test split on one of these datasets using your algorithm. This will allow you to present two datasets that you trained your algorithm on.

3.I did not see that your CNN model could classify the type of ICH. Did the dataset contain the type of hemorrhage in its diagnosis data or in the HGE-Seg files? And if so, can you then provide the accuracy, sensitivity and specificity of your algorithm for each of those individual types of ICH (Epidural, Subdural, Subarachnoid and Intracerebral) ? This will enable the readers (and yourself) to understand if the model is more accurate at correctly diagnosing certain types of ICH than others. see for example (https://doi.org/10.1016/j.heliyon.2024.e30270)

4.This is a follow up to point 3. Please provide confusion matrices for the overall dataset as well as for individual subtypes of ICH (as found in point 3).

5.If you did not perform an analysis such as that in point 3, then your split ratio of train:validate:test probably did not stratify based on the type of ICH. In this case, you would be left with unbalanced classes for these ICH subtypes. Please discuss in the manuscript.

6.Additionally, provide AUC figures for all your 2D, 3D classification models and the image segmentation model.

7.If your 2D classification and segmentation models achieved accuracies of 93.88 and 98.72% respectively, how is it possible that - in some cases- your model correctly segmented the hemorrhagic region while simultaneously identifying the slice as being non-hemorrhagic ? Please explain.

8.I did not quite understand why the bone window classification was performed in the first place, considering its poor performance. Perhaps this needs more explanation in the manuscript.

9.Provide a list of the hyperparameters used by the algorithms as well as any other quantifier that is relevant to the algorithm's prediction capability. Readers must be able to replicate your work.

10.You mention image augmentation techniques to increase the number of images. Please present in a table, the number of images for all the three sets (2d, 3d classfiy and 2d image segment) before and after image augmentation. I would also like to see more images. Present at least three representative images for the 2D and the 3D that were correctly and incorrectly classified as being hemorrhagic or non-hemorrhagic.

11.Did you perform K-fold validation. If not, why not. Please justify.

12.The manuscript needs to be rewritten in third person, past perfect tense. Please correct throughout the paper.

---

1. The two datasets PhysioNet-ICH study and CQ500 dataset were both integrated. Specific references to data collection and preprocessing is between pages 8-12. However the datasets were only integrated for 2D classification models for classifying hemorrhages and subtypes, as hemorrhagic masks were not provided in the CQ500 dataset. Additionally, the RSNA dataset also did not have hemorrhagic masks for CT slices.
2. Once again, see pages 8-12 for full data processing and training, validation and testing
3. Manuscript has changed to add in diagnosing subtypes as well. This was not initially part of the manuscript as it was not the primary focus, but now this classification model has been added as I agree that this provides valuable insights into classification models. Metrics calculated are under the Results and Discussion section, where pages 24-31 show accuracy, sensitivity, and specificity for individual types of ICH. However, because of the integration of multi-label classification, it was decided to remove 3D classification models, as then the paper would be focusing on too many different goals instead of sticking to one model pipeline.
4. Confusion matrices were also provided: see pages 28-31 for individual confusion matrices for each subtype, and page 23 for binary classification confusion matrix
5. Performed classification of subtypes, not necessary. However the difference in the number of subtypes and the impact of each of the metrics (precision, recall, ROC-AUC, and confusion matrices) were discussed in the Results and Discussion section of the paper.
6. AUC figures provided for both 2D classification models, on page 22 for binary classification model, and page 28 for multi-label classification model.
7. This is because only hemorrhagic CT scans were inputted into the 2D image segmentation model, so the model did not segment out hemorrhagic regions while simultaneously identifying the slice as being non-hemorrhagic. This is because of course, non-hemorrhagic slices do not have hemorrhagic masks and couldn't be trained on the 2D U-Net model. See figure 2 on page 4 for the model pipeline.
8. Originally, 3D bone window classification was performed alongside 3D brain classification, to show radiologists which CT scan would be better to diagnose ICH from. However after doing research, it is very clear that a 3D bone model isn't necessary because bone CT scans do not show hemorrhages clearly, and are used mostly for fractures. Hence, this has been removed from the manuscript in addition to the 3D brain classification model
9. Hyperparameters including input image dimensions, normalization, data augmentation, CNN or UNet architecture (filter sizes, kernel size, activation, dropout rate), epochs and batch size for all three models are shown under the Methodology/Models from pages 14-21.
10. Data augmentation techniques, number of images before and after augmentation, and CT scans before and after augmentation are shown on pages 11- 12 for 2D classification, and page 13 for image segmentation model. Additionally, images of correctly and incorrectly classified for the binary classification model are shown in figure 15 on page 24. Images of a subtype (intraparenchymal hemorrhage) incorrectly and correctly classified are shown in figure 20, on page 29.
11. K-Fold validation was not performed on the 2D image segmentation model because data augmentation had already been applied, and the limited dataset increases the risk of overfitting. Repeatedly reusing augmented data during K-fold validation increases this risk of overfitting without providing significant value. K-fold validation for the 2D classification model is not

necessary because the dataset already has over 4000 CT scans, and with augmentation has over 7000 CT scans.

12. Done

---

Thank you for addressing my comments. Accepted. Please do the following:

1.Rewrite the references so that each has the same format. Where there are more than 6 authors, the first 6 authors need to be listed followed by an et al. Each reference must have a live link (DOI) and do NOT use the software automatic numbering to number the references. Instead, please number them manually. Please revise this in the attached document.
2.Include in the attached document, the # of subjects for each category of ICH and include a statement to the effect that these datasets were not balanced before input into the classification algorithm.
When done, please upload the revised document in the communications trail.

---