HealthManagement
Promoting Management and Leadership

Stress-Testing Deep Learning Models: Building Resilient Radiology Models

Deep learning (DL) models have shown promise in automating tasks like bone age prediction, yet concerns exist regarding their readiness in medicine, particularly in radiology, due to their limited robustness to common clinical image variations. A recent study published in Radiology: Artificial Intelligence assessed the robustness of the winning model of the 2017 RSNA Paediatric Bone Age Machine Learning Challenge. authors performed a series of computational "stress tests" to assess the model's performance under various conditions, such as out-of-distribution detection and variable predictions based on user input. Researchers aimed to provide insights into its suitability for real-world clinical use.

Computational Stress Testing of the 16-Bit Model

The study utilised publicly available de-identified images and adapted a framework for dermatology deep learning models to conduct computational stress tests on the 16-bit model, renowned for paediatric bone age prediction. These tests involved various transformations mimicking clinical scenarios, such as rotations, brightness changes, and resolution alterations. Different sets of transformations were applied to two test datasets, RSNA and DHA. The evaluation included spot-checking transformed images to ensure accuracy and comparing the model's performance against radiologist-determined ground truth. The Mann-Whitney U test was used to assess differences between the RSNA and DHA test sets.

Comparative Analysis using Insights from RSNA and DHA Datasets

The study included 1425 images from the RSNA validation set and 1202 images from the DHA set after excluding some due to missing clinical data. The mean bone age for RSNA was 127.2 months, with 54.2% male and 45.8% female patients. For DHA, the mean bone age was 135.3 months, with 51.1% male and 48.9% female patients, representing a diverse population in terms of race, ethnicity, and age. Image transformations were applied to both sets, and 15 out of 21 showed significant differences in mean absolute difference (MAD) between transformed and untransformed images, with consistently higher MADs for transformed images. Horizontal flipping, contrast adjustment, and images with an "L" marker showed no difference in predicted MADs for RSNA but significantly higher MADs for DHA.

Real-world image variations to assess model robustness

The study conducted a comprehensive evaluation of an acclaimed deep-learning model designed for paediatric bone age prediction. While the model demonstrated high performance on unaltered images from both internal (RSNA) and external (DHA) test sets, its robustness to various image transformations mirroring real-world clinical scenarios was scrutinized. These transformations included rotations, brightness and contrast adjustments, flips, pixel inversions, presence of laterality markers, and changes in image resolution. Results revealed that the majority of these transformations led to statistically significant differences in bone age predictions compared to untransformed images. This variability could potentially impact clinical diagnoses, as evidenced by significantly higher rates of clinically significant errors (CSEs) observed with transformed images compared to baseline errors.

Recommendations for Deploying Deep Learning Models in Clinical Settings

Limitations of the study were acknowledged, such as the possibility that the range of image transformations tested may not fully encompass all possible variations encountered in clinical practice. Furthermore, not all transformed images underwent quality control checks, diverging from standard clinical workflows. Additionally, the datasets used were carefully curated, possibly overestimating the model's robustness compared to real-world scenarios with more heterogeneous data.

The study emphasised the importance of rigorous stress testing before deploying deep learning models in clinical settings. Recommendations included augmenting model training with diverse datasets, incorporating mechanisms for out-of-distribution detection, and enforcing image

quality thresholds. Regular quality control checks by radiologists were also encouraged to ensure reliable use of the algorithm in practice.

Overall, the study highlighted the necessity of meticulous stress testing to develop more resilient and clinically applicable deep learning models, ultimately enhancing patient care and safety in medical imaging.

Source: **Radiology: Artificial Intelligence**

Image Credit: **iStock**

Published on : Wed, 24 Apr 2024