SUPPLEMENTARY MATERIAL 5

The Detailed Explanation of Merging Results

We employed the Dice value to judge whether detection results in different layers or different parts of one image belonged to the same fracture, the formula wasusing the formula:

$$\text{Dice} = \frac{2\ \text{area}\ |\ X \cap Y\ |}{\text{area}\ |\ X + Y\ |}$$

where X and Y are rectangular boxes and located in continuous different layers or different parts of one layer. If the Dice value > 0.75, we considered them parts of the same fracture; otherwise, they were considered different fractures. The most frequent category in the sequence was then output as the possible fracture type on the grounds that the less frequent categories might be misidentifications. In the rare event that two categories appeared equally, we selected the one with a higher confidence score as the possible fracture type. Fractures that occupied two layers or fewer were discarded as FPs. The comparison of performance between the structured report with and without discarding fractures present in one or two slices is provided in Table 1. The results showed the model that discarded fractures present in one or two slices had higher precision, higher F1-score, lower number of FPs, and slightly lower recall than the model that did not discard in the three rib fracture types. The development of the software system was carried out in a Linux Ubuntu 18.04 environment (18.04.1 LTS, Bionic Beaver, Boston, MA, USA). Pycharm (2018.1, JetBrains, Praque, Czech Republic) was used as the integrated development environment.

Table 1. Comparison of Performance between Structured Report with and without Discarding One/Two Slices

Metrics	Structured Report Discarded One/Two Slices			Structured Report Reserved One/Two Slices		
	Fresh Fracture	Healing Fracture	Old Fracture	Fresh Fracture	Healing Fracture	Old Fracture
TPs	43 (34–50)	49 (41–55)	38 (29–45)	45 (34–55)	53 (42–62)	40 (30–47)
FNs	2 (0–7)	7 (1–15)	16 (5–27)	0 (0–0)	3 (0–9)	14 (4–24)
FPs	24 (16–34)	12 (5–22)	8 (1–19)	91 (81–102)	40 (30–52)	22 (12–35)
Precision	43/67 = 0.642 (0.507–0.754)	49/61 = 0.803 (0.651–0.915)	38/46 = 0.826 (0.617–0.977)	45/136 = 0.331 (0.250–0.404)	53/93 = 0.570 (0.447–0.667)	40/62 = 0.645 (0.470–0.783)
Recall	43/45 = 0.956 (0.829–1.000)	49/56 = 0.875 (0.741–0.982)	38/54 = 0.704 (0.526–0.898)	45/45 = 1.000 (1.000–1.000)	53/56 = 0.946 (0.830–1.000)	40/54 = 0.741 (0.596–0.922)
F1-score	1.228/1.598 = 0.768 (0.660–0.840)	1.405/1.678 = 0.837 (0.752–0.894)	1.163/1.530 = 0.760 (0.637–0.841)	0.662/1.331 = 0.497 (0.400–0.576)	1.078/1.516 = 0.711 (0.609–0.785)	0.956/1.386 = 0.690 (0.566–0.764)

Corresponding 95% confidence intervals, shown inside parentheses, were estimated by using bootstrapping with 1000 bootstraps and randomly sampled at lesions level. FNs = false negatives, FPs = false positives, TPs = true positives