SUPPLEMENTARY MATERIAL 4

The Detailed Explanation of Indicators

The study detecting rib fractures of one category actually concerned a binary classification. A true positive (TP) occurs when the sample is positive and the prediction is positive. A false positive (FP) occurs when the sample is negative but the prediction is positive. A false negative (FN) occurs when the sample is positive but the prediction is negative. The following is a detailed description of the three indicators we used to evaluate performance.

Precision was determined by how many of the predicted positive samples were TPs.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall was determined by how many positive examples in the real sample were predicted to be positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score was is an indicator used in statistics to measure the performance of classification models. It could be interpreted as the harmonic mean value of accuracy and recall, with a maximum of 1 and a minimum of 0.

$$F_1 = 2 \times \frac{\text{precision x recall}}{\text{precision + recall}}$$

Therefore, the F1-score was adopted in our study to represent the comprehensive performance of the CNN model.

In the following reader test, a TP was defined as correct categorization. A FN was defined as a missed diagnosis or misdiagnosis of some type of fracture. The diagnosis of some types did not match the GT and was defined as a FP of this type. These definitions were applied to both the CNN model and radiologists.