Introduction:
Most of the published papers on AI based diagnosis have focused on the algorithm's diagnostic performance in a 'binary' setting (i.e. disease vs no disease). However, no study evaluated the actual value for the clinicians of an AI based approach in diagnostic. Detection of Traumatic thoracolumbar (TL) fractures is challenging on planar radiographs, resulting in significant rates of missed diagnoses (30-60%), thus constituting a field in which a performance improvement is needed. Aim of this study is therefore to evaluate the value provided by AI generated saliency maps (SM), i.e. the maps that highlight the AI identified region of interests
Methods:
An AI model aimed at identifying TL fractures on plain radiographs was trained and tested on 567 single vertebrae images. Three expert spine surgeons established the Ground Truth (GT) using CT and MRI to confirm the presence of the fracture. From the test set, 12 cases (6 with a GT of fracture, 6 with a GT of no fracture, associated with varying levels of algorithm confidence) were selected and the corresponding SMs were generated and shown to 7 independent evaluators with different grade of experience; the evaluators were requested to: (1) identify the presence or absence of a fracture before and after the saliency map was shown; (2) grade, with a score from 1 (low) to 6 (high) the pertinency (correlation between the map and the human diagnosis) and the utility (the perceived utility in confirming or not the initial diagnosis) of the SM. Furthermore, the usefulness of the SM was evaluated through the rate of correct change in diagnosis after the maps had been shown. Finally, the obtained scores were correlated with the algorithm confidence for the specific case
Results:
Of the selected SM, 8 had an agreement between the AI diagnosis and the GT, while in 4 the diagnosis was discordant. The pertinency of the map was found higher when the AI diagnosis was the same as the GT and the human diagnosis (respectively p-value = .021 and <.000). A positive and significant correlation between the AI confidence score and the perceived utility (Spearman: 27%, p-value=.0-27) was found. Furthermore, evaluator with experience < 5 year found the maps more useful than the experts (z-score=2.004; p-value=.0455). Among the 84 evaluation we found 12 diagnostic errors in respect to the GT, 6 (50%) of which were reverted after the SM evaluation
Discussion:
While the percieved pertinency is higher when the AI and human diagnosis are concordant, the percieved utility correlates with the AI confidence. This highlights the fact that to be considered helpful, the AI must provide not only the diagnosis but also the case specific confidence. Furthermore, the perceived utility was higher among less experienced users and overall the SM were useful in improving the human diagnostic accuracy. This is the first, proof of concept study that show the value of AI generated SM in improving the human diagnostic performance

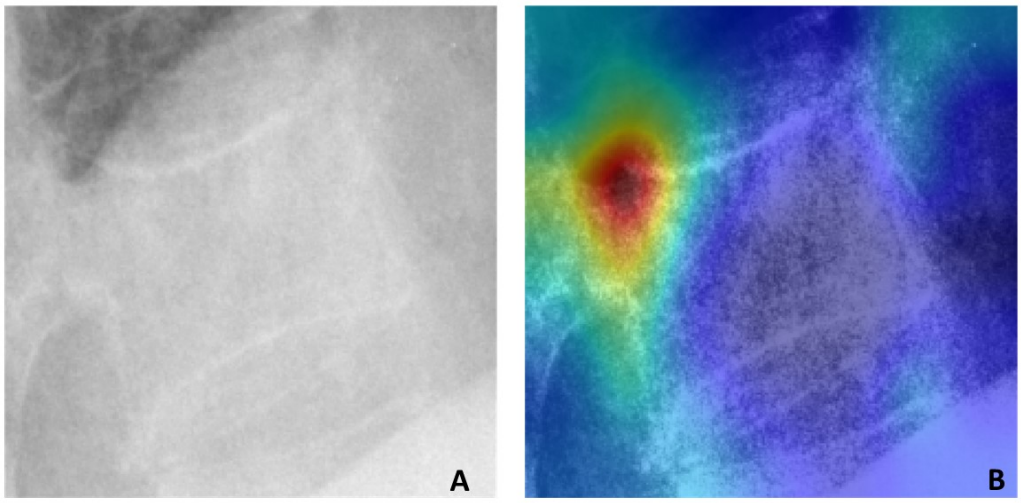

Figure 1. AI generated Saliency Map of case example of a fractured vertebra. The area in red correspond to the Region of Interest identified by the algorithm. In this case, the AI identified a fracture with a confidence score of 99%.