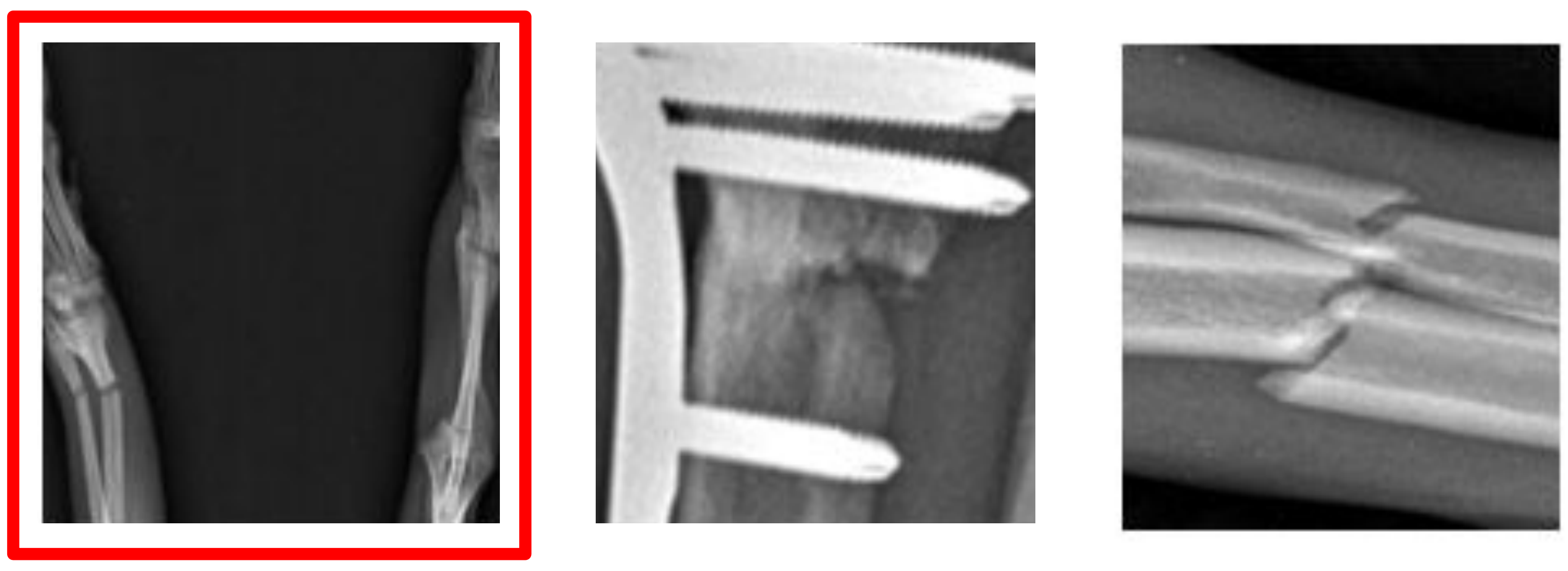
A Saliency-based Clustering Framework for Identifying Aberrant Predictions

Aina Tersol Montserrat
Alexander R. Loftus
Yael Daihes

Not all mistakes are equal. How can we know whether our models are trustworthy?

Imagine a state-of-the-art machine learning model designed to assist veterinary medical practitioners by suggesting whether fractures are present in animal X-rays. Consider a specific scenario where an X-ray shows an unusual pattern caused by a foreign object. In this case, the model might predict the presence of a fracture, even assigning it a high probability score, indicating strong confidence in this diagnosis. However, the model did not make this prediction based on clinical fracture features in the bone region. This would be an aberrant prediction. Note that even if the X-Ray actually has a fracture in it, the prediction is still aberrant if it is based on the wrong input features. [Figure 1] shows an example of aberrant predictions.



a) Aberrant mistake (FP) b) Healing fracture (FP) c) Fracture (TP)

Figure 1. Three different types of predictions, two with false positive predictions (FP) and one with a true positive prediction (TP). Aberrant predictions (a) lower trustworthiness and are undesirable. Logical mistakes (b) are incorrect predictions according to ground truth, but are still reasonable and maintain trustworthiness. (c) is a correct classification.

Methods

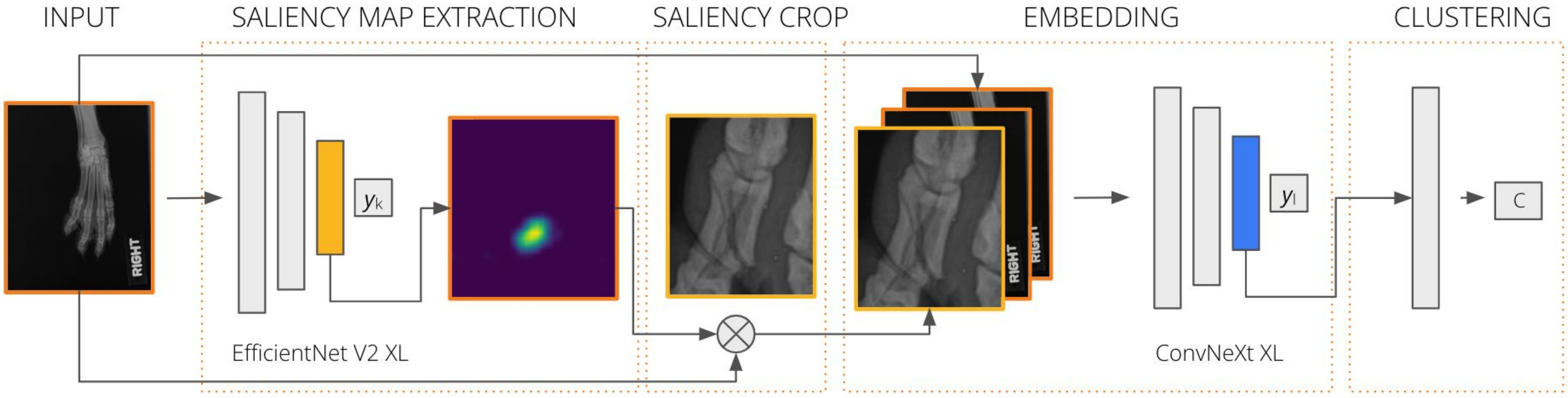


Figure 2. Workflow outlining the steps to identify aberrant predictions. C represents the cluster. y_k and y_l are the classification results.

We first generate saliency maps from the classifier to crop the original radiographic images. These cropped regions are then embedded and subjected to unsupervised clustering with K-nearest neighbors to distinguish between logical and aberrant predictions. This methodology allows us to move beyond simply asking where the fracture is, and instead try to determine whether the highlighted region logically resembles a fractured area.

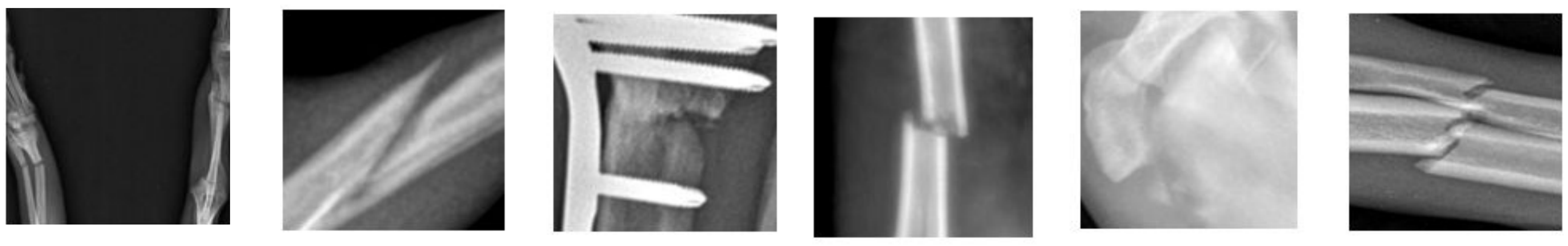
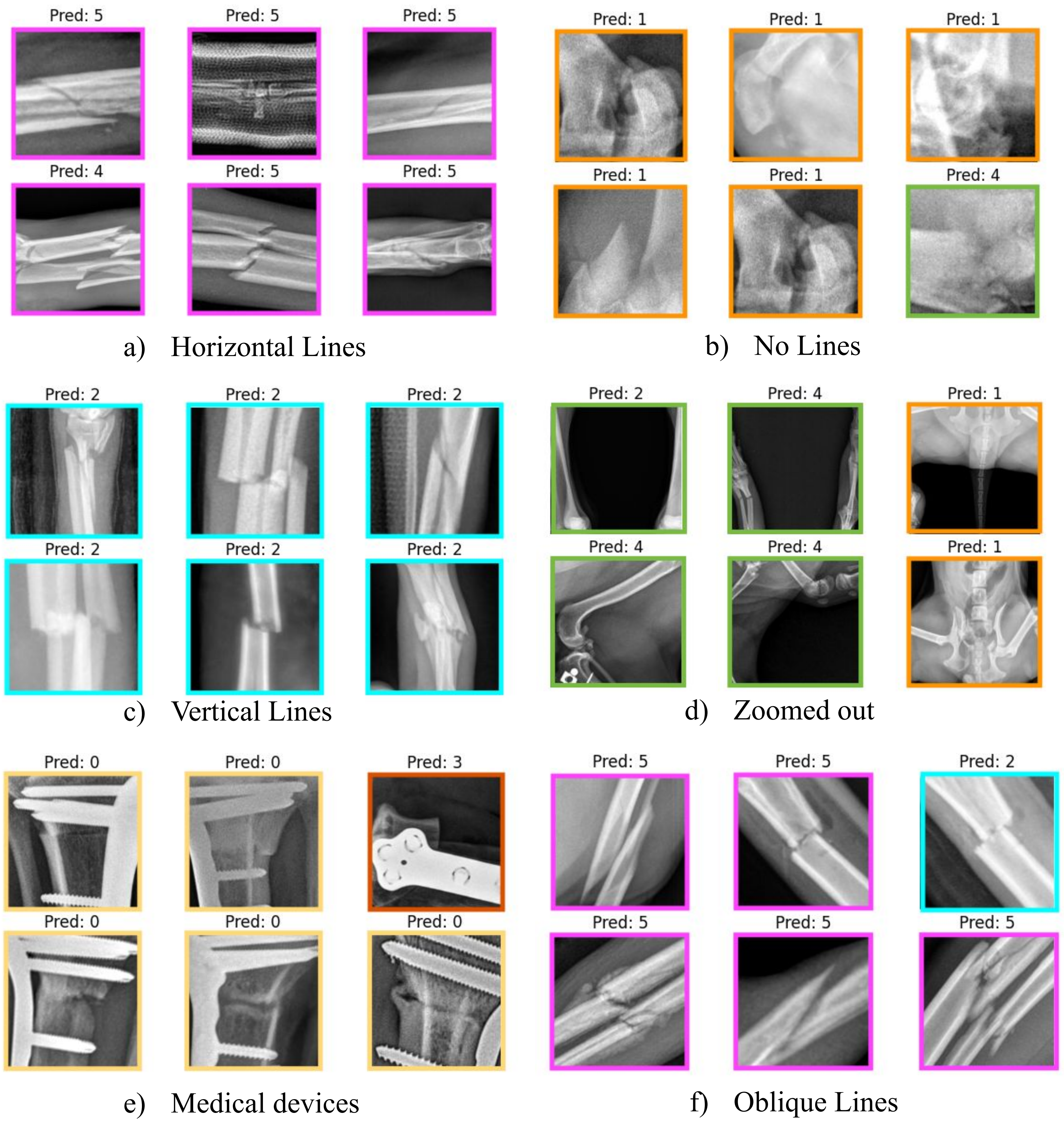


Figure 3. Representative images of saliency map crops. Different visual features can generate a positive fracture classification, with some features (e.g., far left) being obviously aberrant.

Results



a) Horizontal Lines b) No Lines c) Vertical Lines d) Zoomed out e) Medical devices f) Oblique Lines

Figure 4. Performance of the clustering system on images with different visual features. The title and color of each sub-image indicates its final predicted cluster.

We evaluate the performance of our workflow on images with different visual features [Figure 4]. Our findings, illustrated in Figure 5, show that our method localizes aberrant predictions to particular clusters.

Subsequent analysis demonstrates that deleting aberrant clusters improves the original classifier's precision by over 20% [Figure 6], while the recall reduction is a marginal 7.2%.

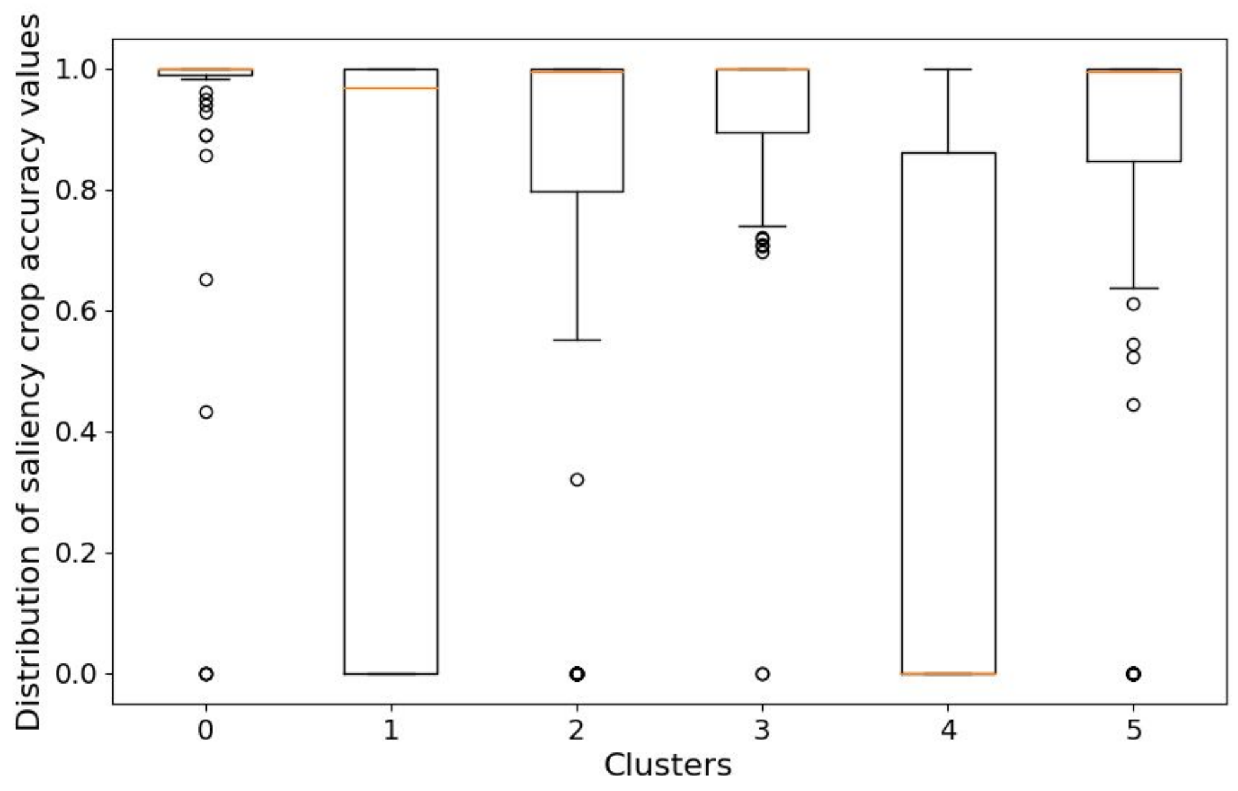


Figure 5. Distribution of saliency map accuracy for each cluster. Clusters 1 and 4 stand out for containing a significant proportion of aberrant predictions

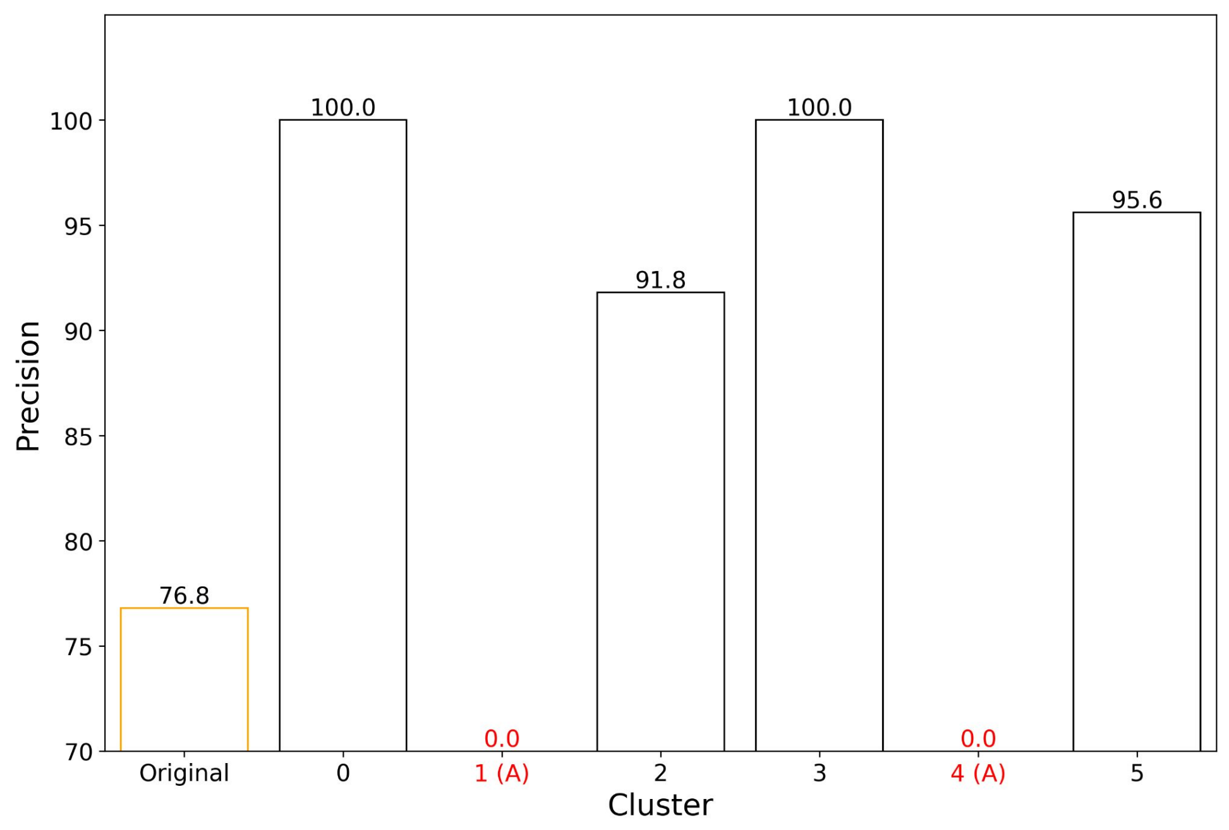


Figure 6. Cluster precision after evaluation in the production model. The precision in clusters 0, 2, 3, and 5 is dramatically better than the original model (76.8%). Clusters 1 and 4 contained mostly aberrant predictions and have a precision of 0, with no true positives.