When Pixels Talk Back:

How Watermarks Disrupt Medical Image Analysis AI

Omkar Kulkarni, Sean Moulton, Anupreet Singh

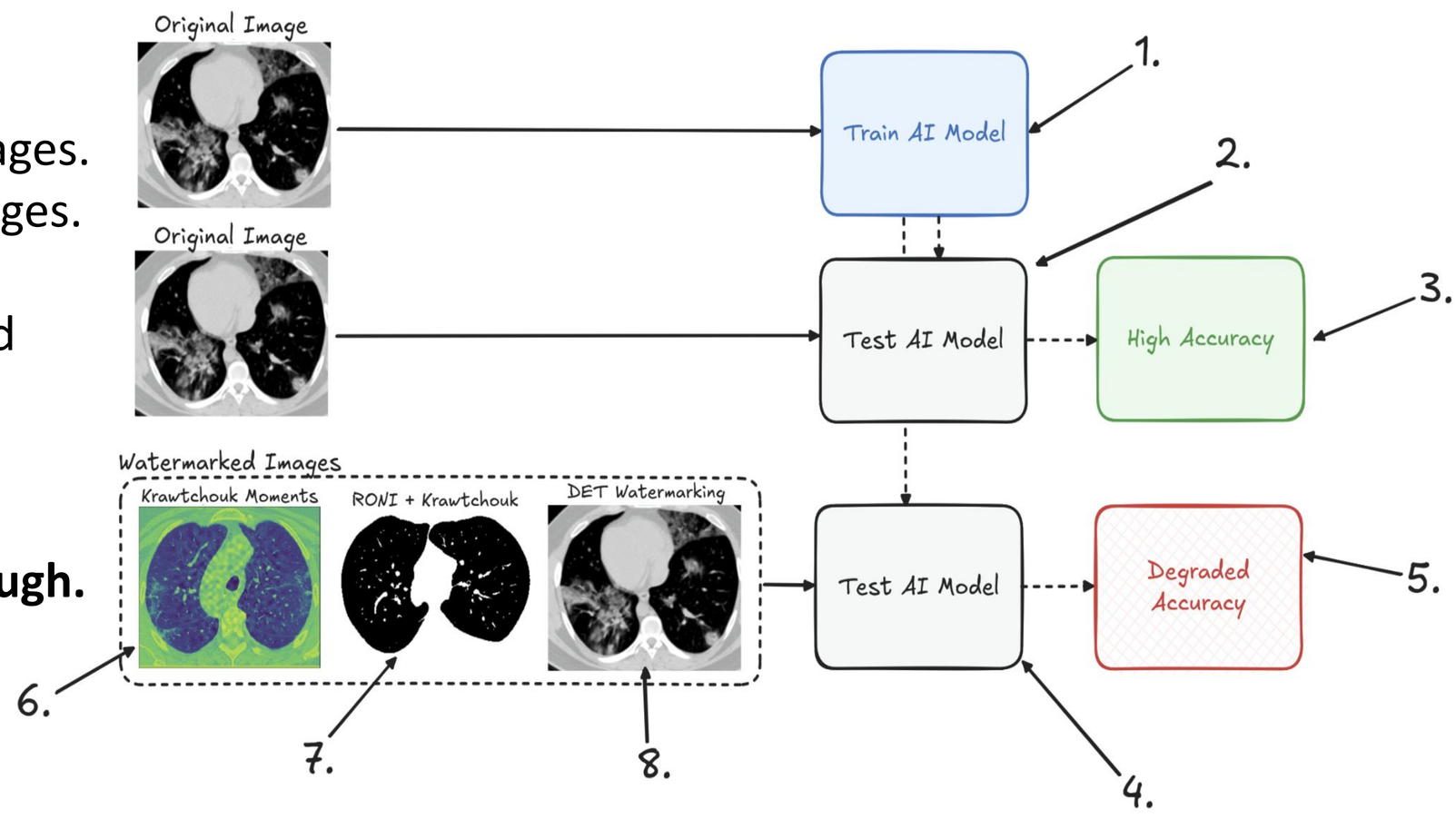
UMBC

Watermarking **degrades the accuracy** of medical image analysis AI.

What can we do about it?

How It Works

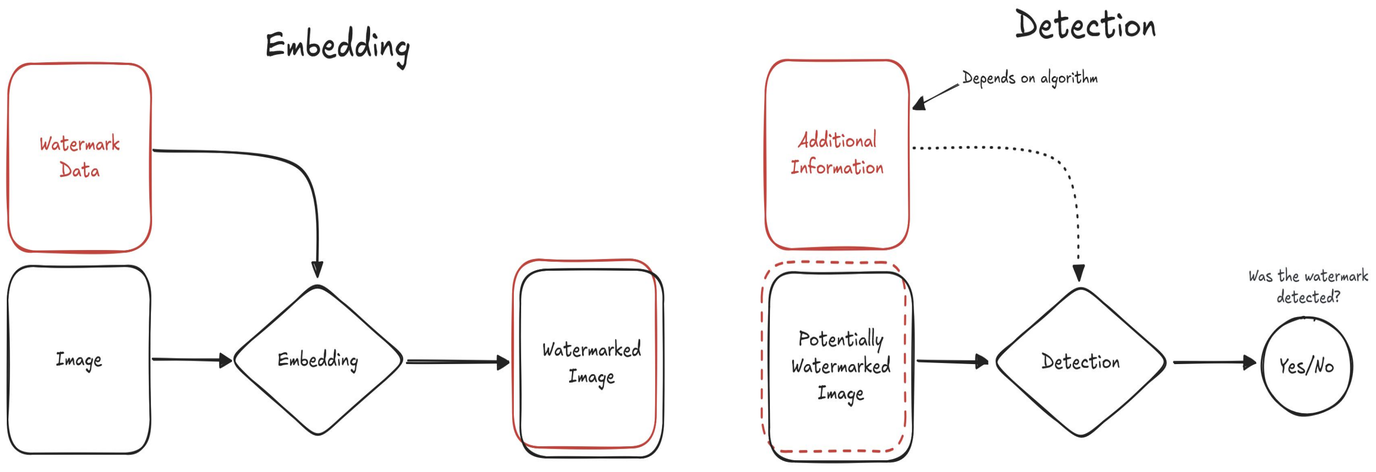
1. Model is trained on the original images.
2. Analysis happens using original images.
3. Good performance.
4. Analysis happens with watermarked data.
5. Degraded performance.
6. **Distortions reduce AI confidence.**
7. **Segmentation helps but is not enough.**
8. **The best performer uses an imperceptible watermark.**



Key Considerations

1. How **visible** is the watermark?
2. How **disruptive** is the watermark?
3. Can I **mitigate** the effects of watermarking?
4. Can I **train** my models better?
5. Can I **use** a different watermarking algorithm?

Watermarking can have knock-on effects for medical image analysis.

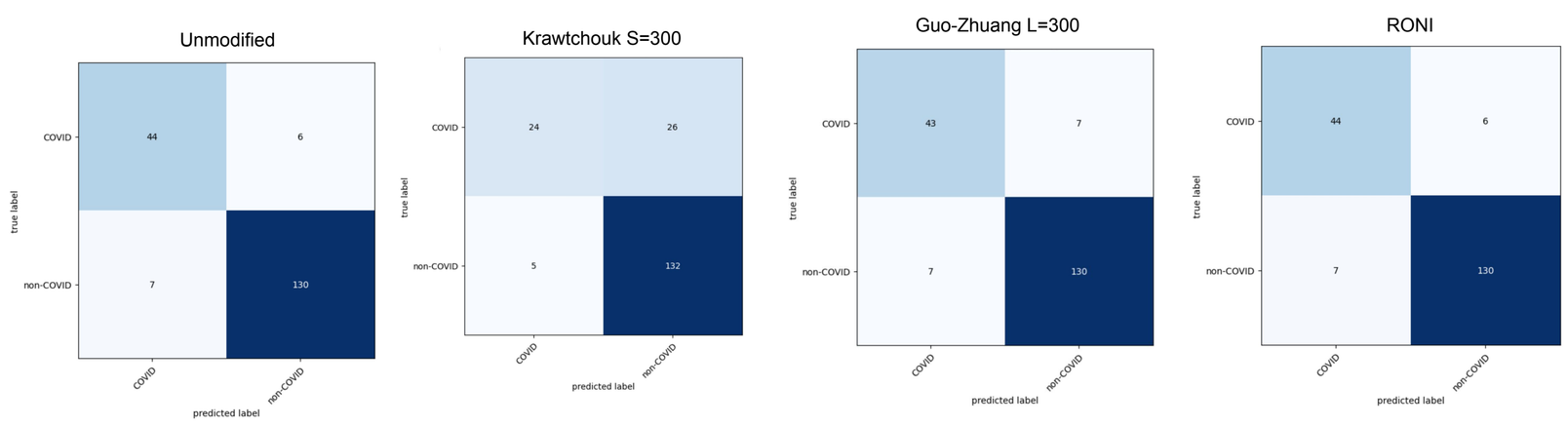


Watermarks help us verify the authenticity and integrity of medical image data. The embedding and detection steps are synonymous with encryption and decryption in secure communication.

Results

1. Classification accuracy drops with structural similarity (SSIM).
2. High rate of false negatives.

False negatives are missed patients.



	Name	Strength	Position	L-Bits	Mean SSIM	Mean PSNR (dB)	Accuracy	Precision	Recall	F1 Score
0	Unaltered	N/A	N/A	N/A	1.000000	∞	0.930481	0.948905	0.955882	0.952381
1	Krawtchouk	50	(0.5, 0.5)	1024	0.988022	42.098871	0.925134	0.941606	0.955556	0.948529
2	Krawtchouk	100	(0.5, 0.5)	1024	0.963324	36.593767	0.898396	0.927007	0.933824	0.930403
3	Krawtchouk	200	(0.5, 0.5)	1024	0.910654	31.217936	0.866310	0.927007	0.894366	0.910394
4	Krawtchouk	300	(0.5, 0.5)	1024	0.868889	28.401042	0.834225	0.963504	0.835443	0.894915
5	Guo-Zhuang	N/A	N/A	75	0.998816	52.394887	0.930481	0.948905	0.955882	0.952381
6	Guo-Zhuang	N/A	N/A	300	0.994847	35.106653	0.925134	0.948905	0.948905	0.948905
7	RONI	N/A	N/A	1024	0.999941	68.406802	0.930481	0.948905	0.955882	0.952381