Explainable Transfer Learning Ensemble AI Model for Pneumothorax Detection

Gábor Orosz MD
Assistant professor, senior lecturer
Semmelweis University
Óbuda University

The Clinical Challenge: Time and Accuracy in the ICU

The Problem

- Critically ill ICU patients: Tension pneumothorax (PTX) is life-threatening within minutes.
- Portable chest X-rays have suboptimal sensitivity.

The Diagnostic Bottleneck

- Lung Ultrasound (LUS): Radiation-free, real-time, bedside.
- The Barrier: Highly operator-dependent and subjective.
- The Goal: Eliminate subjectivity and diagnostic variability.

Research Gaps & Aims

Data Quality

Previous models relied on simulations or animal models.
Lack of diverse human pathology

Validation Gap

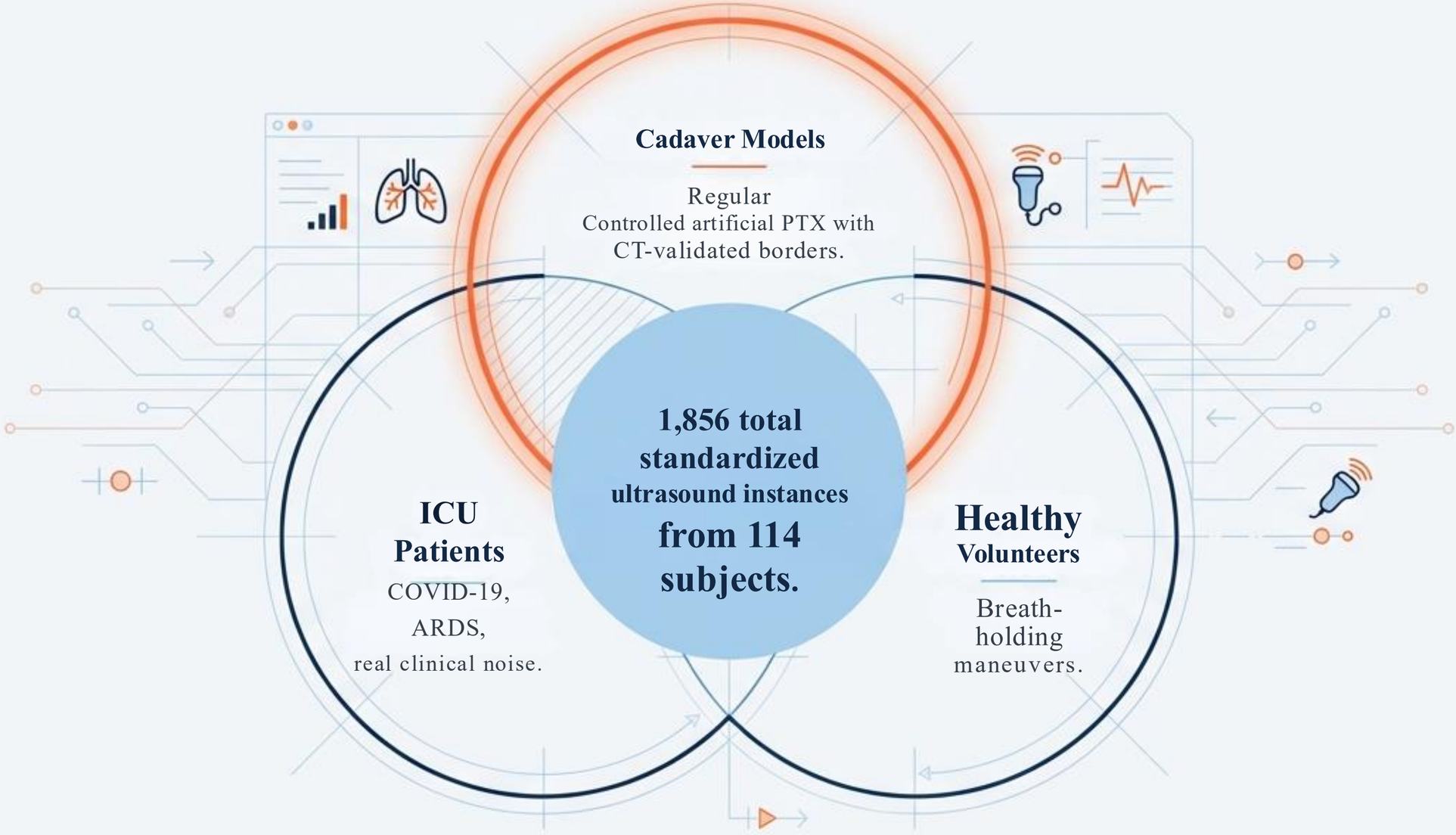
Scarcity of direct benchmarking against highly experienced human clinicians

The Black Box

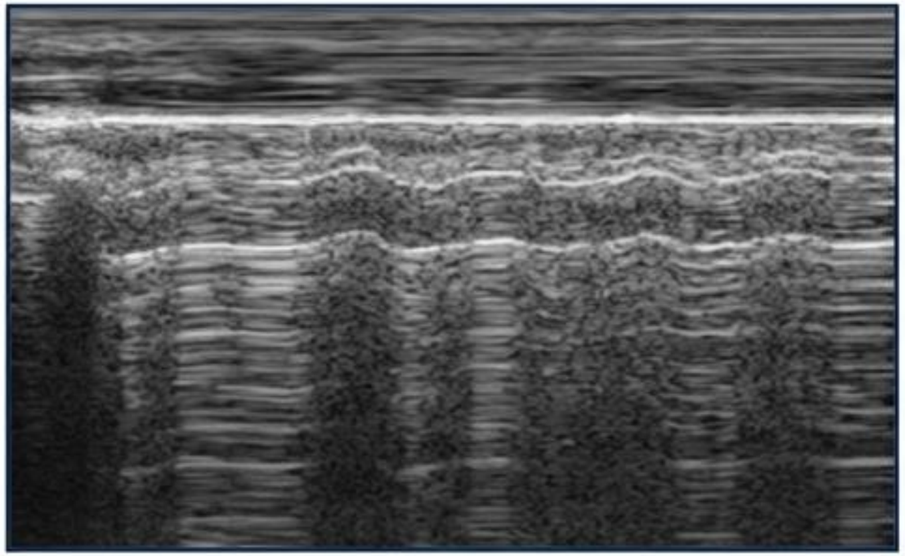
Lack of interpretability prevents clinical trust and adoption

Aim: To develop and validate a robust, explainable AI ensemble trained on a diverse, real-world dataset to match or exceed an expert medical committee.

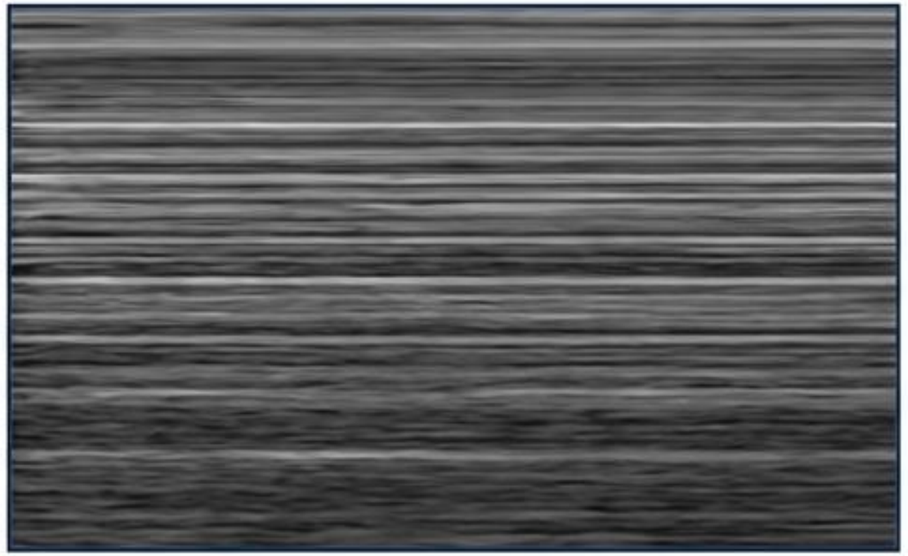
Innovative Dataset Creation: The Golden Standard



Interpreting Diagnostic Signs: B-Mode vs. M-Mode



Normal (Seashore sign) - Granular pattern indicating healthy lung sliding.



PTX (Barcode sign) - Static horizontal lines indicating absent lung sliding.

M-Mode (Motion)

Provides high temporal resolution of pleural movement, transforming dynamic video into a single 2D image for CNN analysis.

State	Sign	Description
Normal Lung	Seashore sign	Granular pattern from healthy pleural sliding.
Pneumothorax	Barcode sign	Static horizontal lines indicating absence of sliding.
The Trap	Lung pulse	Subtle cardiac rhythmic movements. Visually deceptive, NOT a PTX.

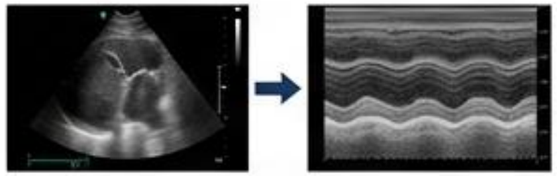
Study Workflow and Pipeline

Step 1: Standardized Data Acquisition

(B-mode clips from ICU, cadavers, volunteers)

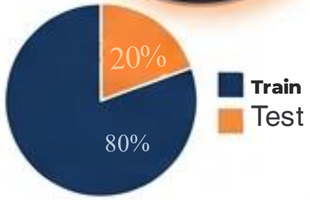
Step 2: Processing

(Automated conversion of B-mode videos to static M-mode)



B-mode ultrasound **Video**

Static M-mode

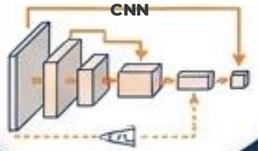


Step 3: Annotation & Split

(Subject-level splits to prevent leakage: 80% Train / 20% Test)

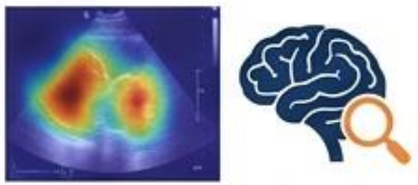
Step 4: AI Training

(Transfer learning and fine-tuning ensemble)



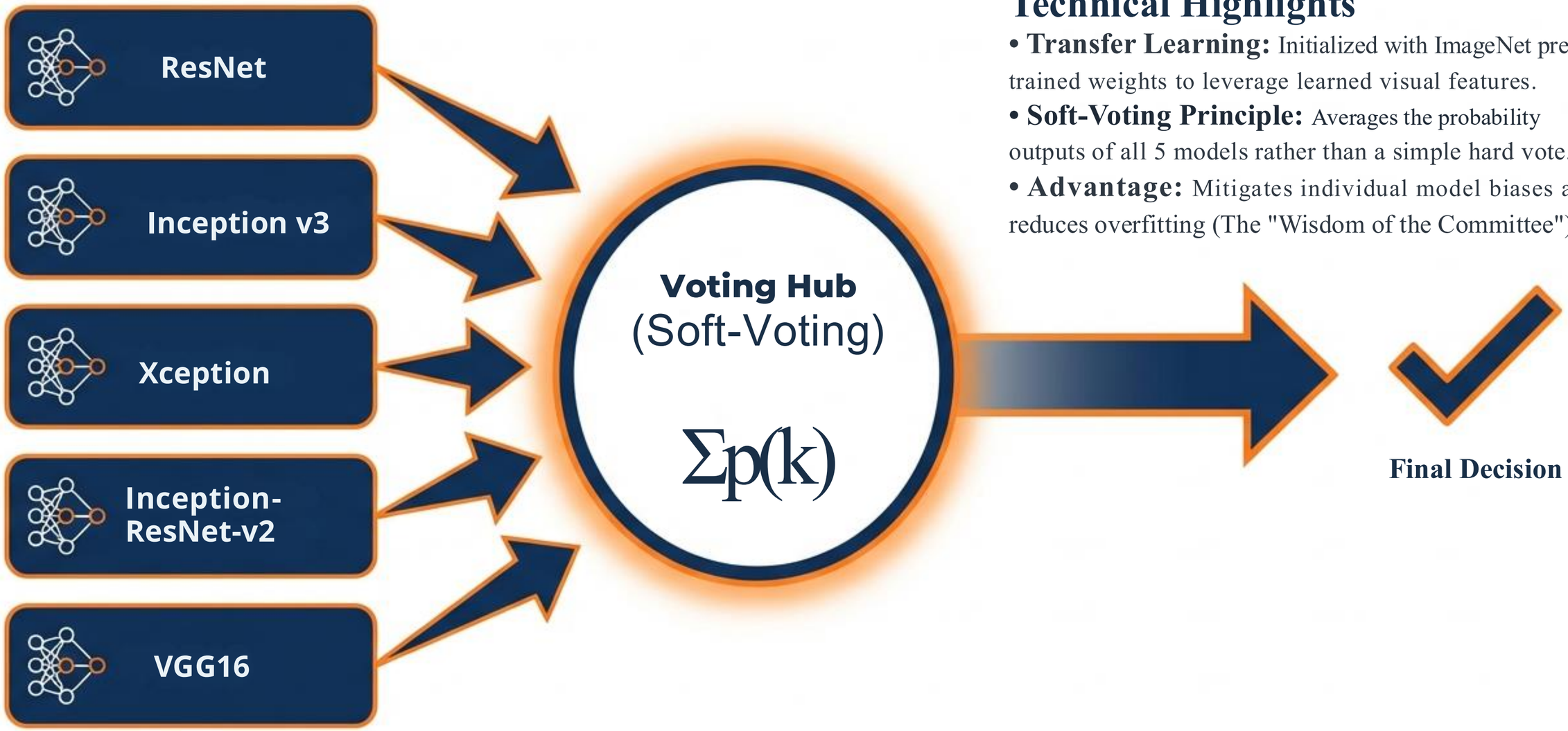
Step 5: Validation

(AI vs. 11 Experts benchmark and Grad-CAM++ explainability)



AI vs. Experts Benchmark & Grad-CAM++ Explainability

Methodology: The Ensemble AI Architecture



Technical Highlights

- **Transfer Learning:** Initialized with ImageNet pre-trained weights to leverage learned visual features.
- **Soft-Voting Principle:** Averages the probability outputs of all 5 models rather than a simple hard vote.
- **Advantage:** Mitigates individual model biases and reduces overfitting (The "Wisdom of the Committee").

The Benchmark Setup: AI vs. Human Expert Panel

Vs

The Human Panel

- 11 POCUS Physicians.
- >10 years of clinical ultrasound experience each.
- Evaluated using both B-mode (video) and M-mode (static).

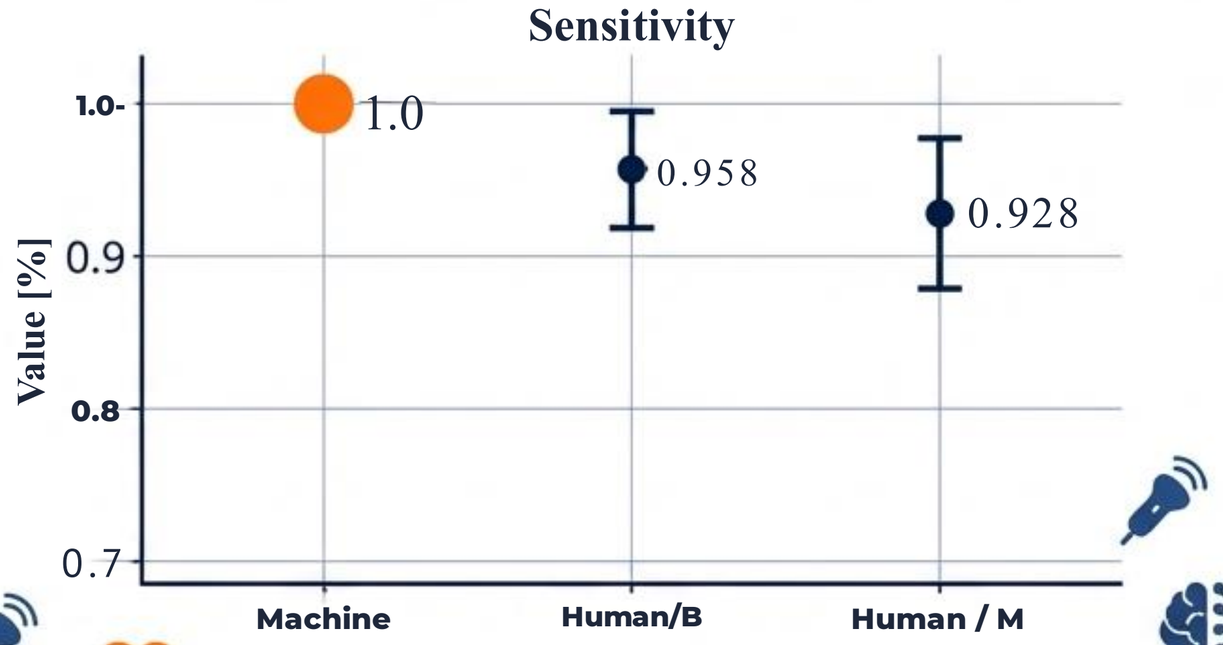
The Blinded Test Set

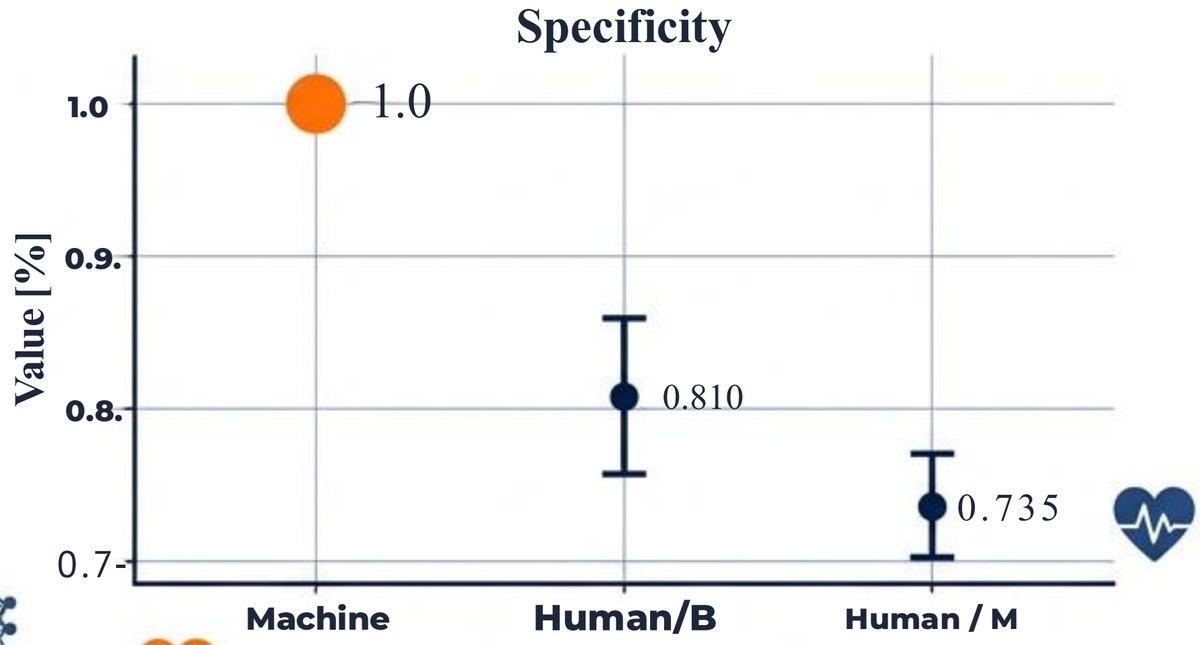
- 48 independent cases (never seen by AI in training).

Perfectly balanced: 24 PTX (Positive) / 24 Non-PTX (Negative).

- Included intentionally difficult breath-holding cases.

Primary Results: The AI as the Idealized Expert





Metric	AI Model (M-mode)	Human Average (B-mode)	Human Average (M-mode)
Sensitivity	100.0%	95.8%	92.8%
Specificity	100.0%	81.0%	73.5%

The AI achieved perfect diagnostic performance, effectively matching the combined consensus of an 11-expert panel.

The Specificity Advantage

The Specificity Advantage:

- AI vs. Human Specificity difference is statistically significant ($p < 0.05$).
- Human experts tend to over-diagnose, resulting in high false-positive rates and unnecessary chest tubes.

M-Mode Capability:

- Humans struggled with M-mode specificity, dropping to 73.5%.
- The AI leveraged M-mode with 100% accuracy, proving the format is information-rich if interpreted correctly.

Bottom Line: The AI acts as an idealized expert-catching all true pneumothoraces without raising false alarms.

Analyzing Human Error: Inter-Observer Variability

	B-mode Sensitivity	B-mode Specificity	M-mode Sensitivity	M-mode Specificity
expert #1	1.00	0.88	1.00	0.67
expert #2	0.96	0.54	0.96	0.46
expert #3	0.83	1.00	0.83	1.00
expert #4	1.00	0.79	0.96	0.75
expert #5	1.00	0.88	0.96	0.88
expert #6	0.92	0.96	0.92	0.88
expert #7	1.00	0.75	0.79	0.50
expert #8	1.00	0.62	1.00	0.54
expert #9	0.96	0.79	0.96	0.79
expert #10	0.88	0.88	0.92	0.67
expert #11	1.00	0.83	0.92	0.96
AI model	1.00	1.00	1.00	1.00

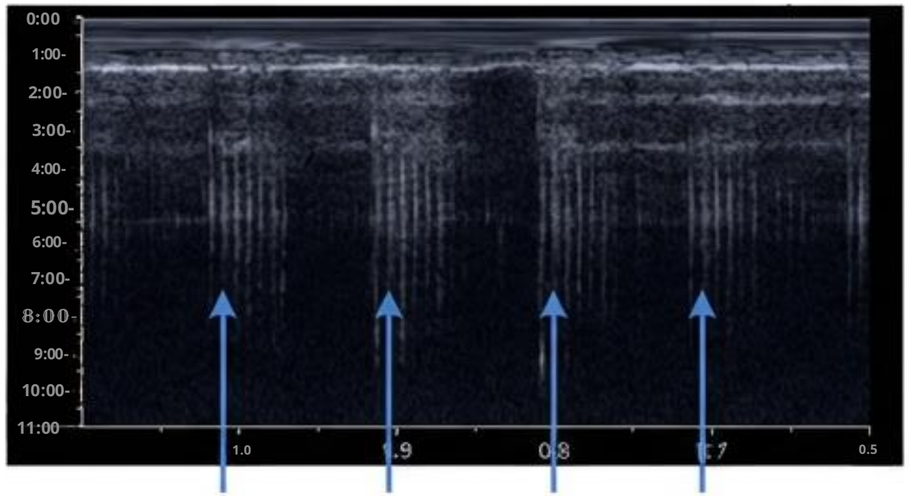
Human Variability **and Uncertainty**

AI Consistency (100%)

- **The Human Trade-Off:** High sensitivity experts often sacrificed specificity (being overly cautious), while high specificity experts missed true pneumothoraces.
- **AI Consistency:** The AI operates at the ideal 100/100 point, eliminating inter-observer variability and standardizing the diagnosis.

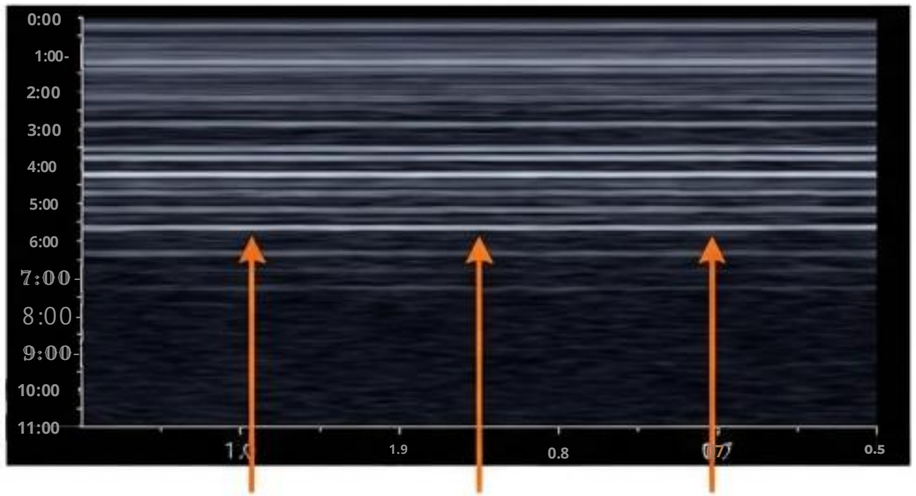
Analyzing Human Error: The Trap Cases

Lung Pulse (Breath-holding)



Subtle rhythmic cardiac oscillations (T-lines)

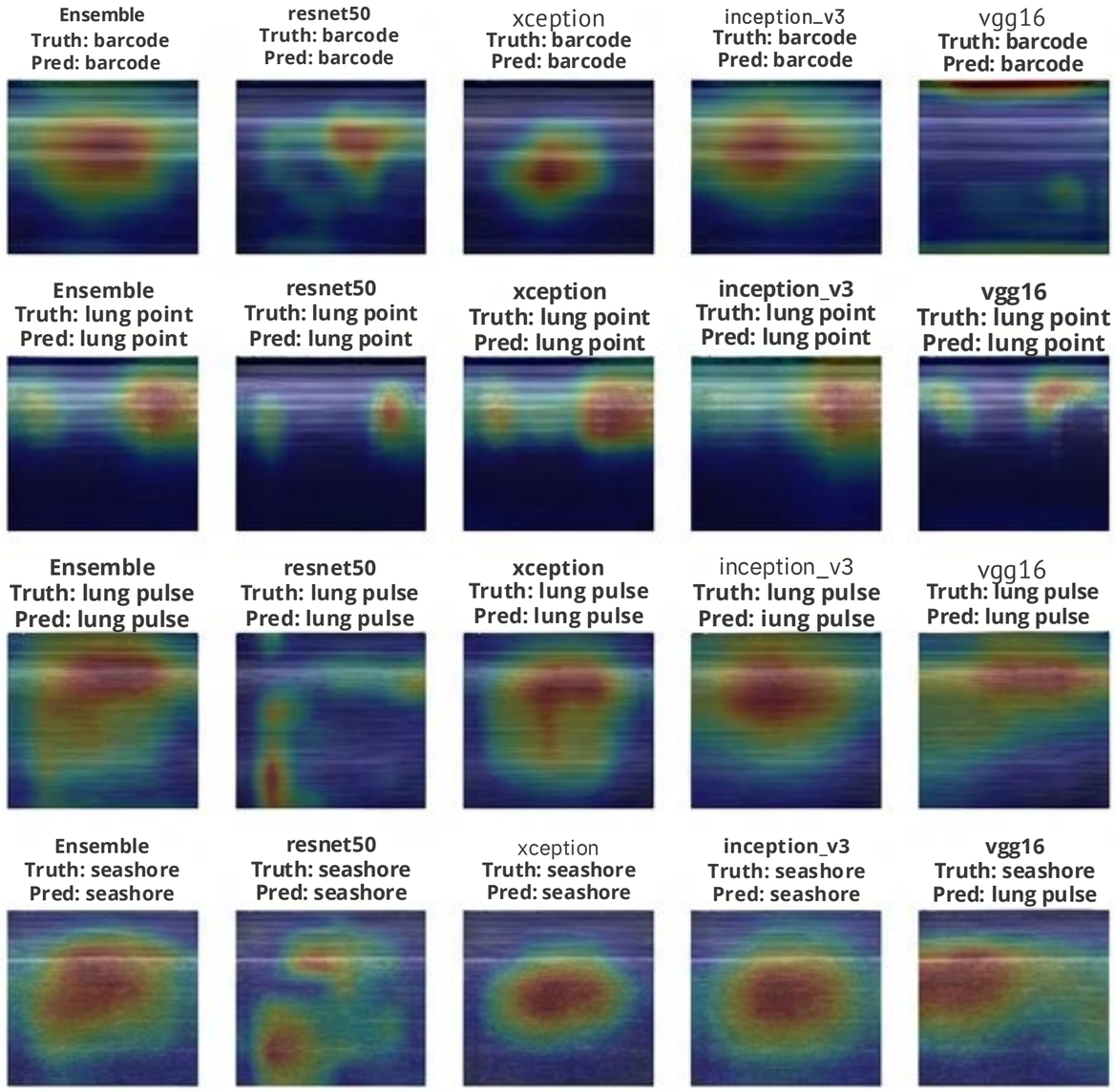
Pneumothorax (Barcode)



Complete immobility (Barcode sign)

- **The Trap:** The 5 most challenging test cases were breath-holding volunteers.
  The Human Error: Experts frequently misread 'Lung Pulse' as the static 'Barcode' sign of PTX, causing high false-positive rates.
  The AI Victory: The ensemble model successfully learned to differentiate subtle cardiac rhythms from true pneumothorax, maintaining 100% accuracy.

Shattering the Black Box: Grad-CAM++ Explainability

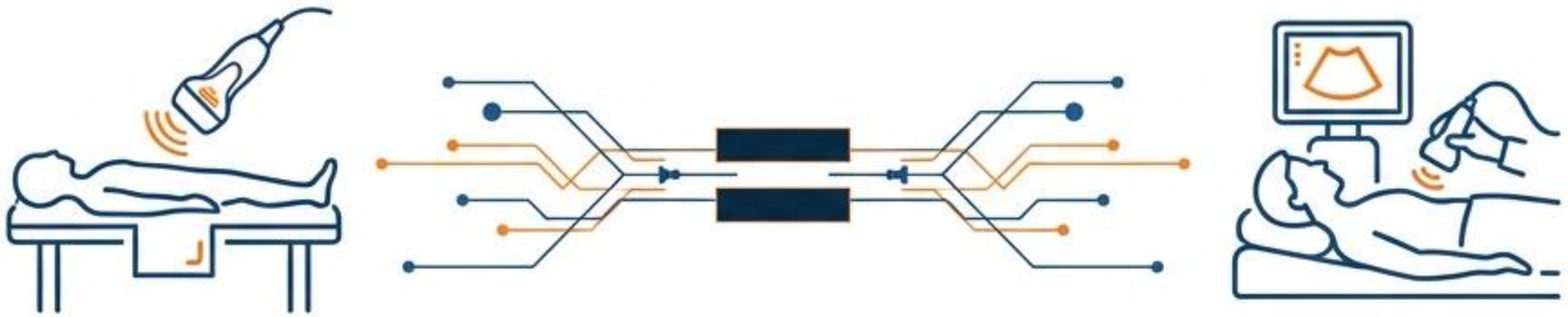


- **Grad-CAM++ Explainability:** Visualizes the exact pixel regions that drove the Al's classification. Heatmaps consistently focused on the crucial pleural line and relevant artifact patterns.
- **Clinical Validation:** 5 human experts blinded to the Al's final diagnosis evaluated the heatmaps.
- **The Result:** 84.2% of heatmaps were rated as anatomically and clinically relevant, building crucial physician trust.

Generalizability: From Cadavers to Live Patients

Cadaver Models

Live Patients



Sensitivity: 100.0%

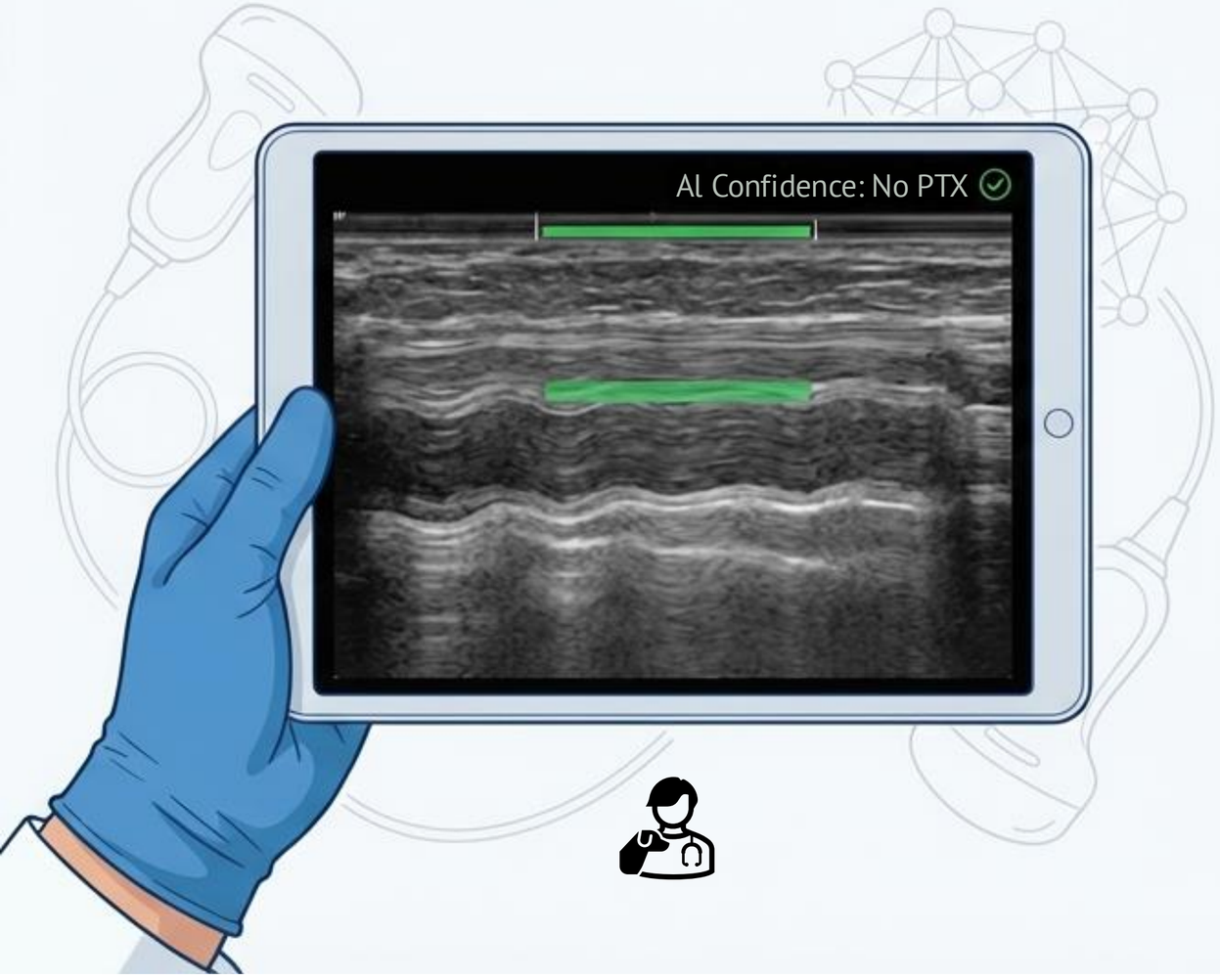
Sensitivity: 100.0%

- **The Translation Question:** Can an AI trained heavily on artificially induced cadaver PTX work on real, noisy ICU patients?

The Answer: Yes. Statistical analysis showed no significant difference between the sources.

- **Conclusion:** The AI successfully transferred its knowledge from controlled environments to chaotic clinical reality.

Clinical Relevance & Future Outlook



The Second Reader: Acts as an automated safety net to support tired residents and standardizes objective decisions regardless of operator experience.

Patient Safety: Drastic reduction in false positives equates to fewer unnecessary, invasive chest tube insertions.

Future Outlook: Model is primed for real-time, bedside integration. Next steps include large-scale multicenter clinical trials to validate global generalizability.