The intelligent stethoscope of the 21st century: clinical workflow standardization, realistic training model development, and a bedside AI-based diagnostic assistant for lung ultrasound in the critically ill patients

PhD thesis

Gábor Orosz MD

Division of Health Sciences
Semmelweis University

Supervisor: Tamás Haidegger, PhD, habil.

Official reviewers: Róbert Gyula Almási MD, PhD, habil.
Vladimir László Tadity, PhD

Head of the Complex Examination Committee:
Katalin Darvas MD, PhD, habil.

Members of the Final Examination Committee:
Tamás Ferenci, PhD, habil.
Enikő Kovács MD, PhD

Budapest, 2026

1. Introduction

When *Laennec* introduced the stethoscope in *1816*, its significance was not the wooden tube itself, but the fact that listening to the chest became a structured, teachable and repeatable part of the physical examination. Lung-focused point-of-care ultrasound (*POCUS*) occupies the same position today: a portable, radiation-free extension of the bedside examination that shows the chest in real time, at the moment the decision has to be made. An ultrasound system coupled with an explainable artificial intelligence (*xAI*) model—benchmarked against expert consensus and able to show which image region drove its output—is no longer only a probe; it functions as an *"intelligent stethoscope"*.

The bedside workflow already provides *immediacy*, *repeatability*, and *freedom from radiation and patient transport*. What it has not provided is *reproducibility between operators*, *standardized acquisition conditions*, and a *traceable link from examination to interpretation*—or an independent second reading. ***None of these limitations is resolved by improving diagnostic accuracy alone.*** **These define the objectives of this doctoral thesis, which treats the intelligent stethoscope not as a piece of hardware but as a complete bedside pathway: from protocol standardization and training-model development, through explainable machine learning, to a deployable clinician-in-the-loop bedside system.**

2. Objectives

Five gaps have been identified in the literature defined by five connected studies; each coupled with a testable hypothesis. The dependency between the studies is methodological rather than merely chronological: each study supplied a prerequisite for the next one.

Objective 1. *AI models for lung ultrasound are typically trained on single-center data and validated against a single-reader reference—or no human comparator at all.* **The aim** was to develop and validate an explainable, soft-voting ensemble of five convolutional neural networks for pneumothorax detection on *M-mode* images, trained on clinical, cadaveric, and volunteer data, benchmarked against 11 experienced clinicians, and made interpretable with Gradient-weighted class activation map (*Grad-CAM++*) heatmaps. **Hypothesis:** expert-level sensitivity with superior specificity can be reached, and explanations corresponding to anatomically meaningful regions gained.

Objective 2. *Comprehensive 12- to 14-region scoring protocols are effective but costly in time and staff; no validated lighter, single-operator alternative existed for ventilated patients.* **The aim** was to assess the feasibility and reliability of the *BLUE-LUSS* protocol—the three *BLUE* points per hemithorax fused with semi-quantitative lung ultrasound scoring (*cLUSS*, *qLUSS*)—at mechanically ventilated *COVID-19* patients. **Hypothesis:** excellent inter-rater reliability of summarized scores and inverse correlation with the PaO_2/FiO_2 ratio can be reached.

Objective 3. *Automated lung ultrasound analysis is predominantly single-task and validated at image level; multi-class semantic mapping and real-time integration remain uncommon.* **The aim** was to automate image processing with a *U-Net*-based multi-class semantic segmentation network validated on unseen patients, progressing toward real-time, bedside-capable processing. **Hypothesis:** mean validation accuracy above *85% would* provide a foundation for automated scoring and protocol-level decision support.

Objective 4. *Training and validation models rarely reproduce complex artifacts or dynamic physiology and no labelled clinical ultrasound dataset of pneumothorax exists, because the*

condition is rare, transient, and urgent. **The aim** was to develop and validate a minimally invasive human cadaveric pneumothorax model—controlled intrapleural air insufflation while the thorax being left intact—serving both ultrasound training and *AI* ground-truth generation. **Hypothesis:** reliable reproduction of the key sonographic signs with substantial expert agreement on quality and educational utility can be provided.

Objective 5. *No widely adopted acquisition-to-interpretation framework exists for AI-assisted lung POCUS; deployment remains an open question.* **The aim** was to propose a comprehensive framework—standardized acquisition settings, dataset diversity, and expert benchmarking requirements, explainability on recognizable sonographic landmarks, preserved clinician oversight—and to realize it as a deployable, clinician-in-the-loop bedside system. **Hypothesis:** a unified, yet generic framework can be provided.

3. Methods

3.1. Study populations and data assets

My research is composed of human clinical, ex vivo and in silico studies as well. For the human studies, all cohorts were collected prospectively at *Semmelweis University* during and after the *COVID-19* pandemic (*2020–2024*), with partitioning at subject level throughout: all material from a given patient, volunteer or cadaver was assigned exclusively to training or to testing. The clinical *AI* cohort comprised *114* unique subjects yielding *1,856* curated *M-mode* images—positives: *23* cadavers with *CT*-verified induced pneumothorax and *2* living patients with confirmed spontaneous pneumothorax; negatives: *79* living patients with pneumothorax excluded by *CT*, serial chest *X-rays,* and consensus review, and *10* healthy volunteers performing breath-holds (absent sliding with preserved lung pulse—the

classic pitfall). The *BLUE-LUSS* cohort comprised *24* consecutive *SARS-CoV-2 RT-PCR*-positive, mechanically ventilated patients (*December 2021 – April 2022*), *20* analyzed, providing *400* loops. The segmentation dataset comprised *1,097* annotated frames from *40* patients (*22 COVID-19, 18 non-COVID*); the expert benchmark set, *48* balanced cases assessed by *11* blinded clinicians. All studies were approved by the *Semmelweis University Regional and Institutional Committee of Science and Research Ethics* (***No. 303/2021***) and complied with the *Declaration of Helsinki* and *GDPR*, with irreversible pseudo-anonymization; volunteers gave written informed consent.

3.2. Standardized image acquisition

Three ultrasound systems from two manufacturers were used deliberately—*Philips CX50, GE Venue GO* and *GE LOGIQ V2*, convex transducers (*1–5.7 MHz*)—so that neither protocols nor models would be tied to one vendor's image processing method. Settings were fixed across every study: depth at the pleural line plus *3–8* cm; single focal zone at the pleura, multifocus off; gain optimized for the primary artifacts; tissue harmonic imaging, *XRes, CrossXBeam,* and *SRI* disabled, preserving native artifact patterns; mechanical index < *0.7*, soft-tissue thermal index < *1.0*. Examinations followed the international protocols; *4–6*-second loops (clinical studies) and *10*-second clips (cadaveric study) were stored in *DICOM* (Digital imaging and communications in medicine) format with pseudo-anonymization at capture and encrypted storage.

3.3. The human cadaveric pneumothorax model

Five fresh, unembalmed cadavers (*3* females aged *79–88*, *2* males aged *70–74*) with intact thoracic anatomy served for the basis of the feasibility study. After intubation, ventilation was

simulated with a resuscitation bag and *PEEP* (positive end-expiratory pressure) valve (*5* cmH_2O, *6–7* mL/kg, *12*/min). A *3* mm cannula was inserted at the safety triangle (fifth intercostal space, mid-axillary line) by blunt dissection, and pneumothorax was induced with a membrane pump (up to *140* cmH_2O), titratable and reversible by unclamping. A second, evolved cohort of *23* mechanically ventilated cadavers (*Newport HT70*) provided the *AI* ground truth: pneumothorax was induced at *200, 400,* and *600* mL, the sonographic pleural borders were marked with permanent ink and metallic pins, verified by on-site *CT* reviewed by a blinded radiologist, and only then were images acquired. The library included *141* expert-evaluated clips and *826 CT*-verified *M-mode* images.

3.4. The BLUE-LUSS protocol

The *BLUE-LUSS* integrates the three *BLUE* points per hemithorax (upper, lower, *PLAPS*) with two semi-quantitative scores: *cLUSS* (*0*: A-lines or at most two B-lines; *1*: ≥3 well-spaced B-lines; *2*: coalescent B-lines; *3*: tissue-like pattern) and *qLUSS* (percentage of involved pleura). Each point was recorded in longitudinal and transverse orientation by a single operator, without repositioning or extra staff. The *400* loops were scored offline by four independent intensivists (each > *7* years of *LUS* experience), blinded to each other and to clinical data. The study was powered a priori ($\alpha = 0.05$, power = *0.8*, *ICC 0.8 ± 0.15*, n = *20–24*); scores were correlated with PaO_2/FiO_2, inflammatory biomarkers (*WBC, CRP, PCT*) and severity scores (*APACHE II, CURB-65*).

3.5. AI model development and bedside deployment

The central methodological step converts a dynamic pleural-motion problem into a static classification task: a fixed vertical scanline through the pleural line is extracted across sequential

B-mode frames and concatenated into a *256 × 256 M-mode* image, with no aggressive denoising or enhancement. The classifier is a soft-voting ensemble of five convolutional neural networks—*CNN*s (*ResNet50, Inception v3, Xception, Inception-ResNet-v2, VGG16*), *ImageNet*-initialized and fully fine-tuned, averaging probabilities over four classes (normal sliding, barcode sign, lung point, lung pulse). The *1,856* images were split *80/20* at subject level (*1,485* training, *371* hold-out); five-fold cross-validation set the hyperparameters (batch *32*, learning rate 10^{-4} with cosine annealing, *Adam* optimizer, *30* epochs, early stopping). *Grad-CAM++*, averaged across the five models, provided visual explanations; *48* heatmaps were rated binary for clinical relevance by five independent, blinded experts. Benchmarking used a balanced *48*-case subset (*24* pneumothorax, *24* not) assessed independently in both *B-mode* and *M-mode* by *11* clinicians from five teaching hospitals, each with > *10* years of *LUS* experience.

For automation, a *U-Net* encoder–decoder with skip connections and *16* base filters performed multi-class segmentation (thoracic wall, ribs, rib shadows, B-lines, septal rockets, abnormal pleural line, among others), annotated in *3D Slicer* software on every fifth frame of a full respiratory cycle and validated with leave-one-out cross-validation at patient level. For bedside deployment, an edge runtime was built on a *Raspberry Pi 5* single board computer with *Epiphan AV.io* video capture: *YOLOv8n*-based pleural localization, consensus scanline, *M-mode* generation, ensemble classification and on-demand *Grad-CAM++*. A hardware-accelerated branch compiled the detector and the model ensemble to *HAILO HEF* (Hailo execution framework) assets (explainability host-side); a native touchscreen dashboard, an automation variant with fail-safe rules, and a structured feedback and provenance layer completed the system, archived on *Zenodo* (DOI: 10.5281/zenodo.19541650).

3.6. Statistical analysis

Diagnostic metrics were reported with exact *Clopper–Pearson 95*% confidence intervals. The *AI*–human comparison used mixed-effects logistic regression with random intercepts for cases, modes, and subjects, three model configurations (*M-mode* only, both modes, evaluator × mode interaction) and likelihood-ratio χ^2 tests (two-tailed, $\alpha = 0.05$). Reliability used intraclass correlation coefficients - *ICC*s (two-way random-effects, absolute agreement), *Cohen's* κ with the *Fleiss–Cuzick* extension for loop scoring, and *Fleiss'* κ for multi-rater ratings; correlations used *Pearson's or Spearman's* coefficients. Analyses ran in *R 4.5.0* (epiR, lme4, IRR) and *StatsDirect 3.1.20*.

4. Results

4.1 Objective 1: explainable AI ensemble and its expert benchmark

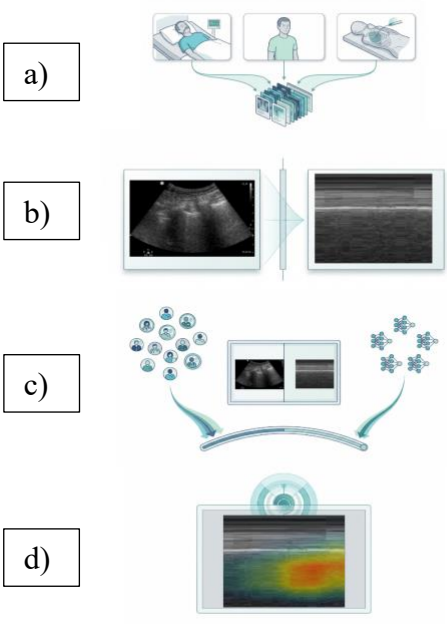

Figure 1: Workflow of Objective 1—development, expert benchmarking, and explainability validation of the M-mode ensemble model for pneumothorax detection. (a) Three complementary data sources feed a pooled dataset of *1,856 M-mode* images: *ICU* (intensive care unit) patients with heterogeneous respiratory pathology, *healthy volunteers* performing breath-holds, and *human cadaveric models* with controlled, *CT*-verified pneumothorax—deliberate diversification of pleural motion patterns, not merely more samples. (b) A scanline perpendicular to the pleural line converts the dynamic *B-mode* clip into a static *M-mode* image, the classifier's input. (c) Head-to-head evaluation on a balanced *48*-case subset not used in training: 1*1* expert clinicians versus the soft-voting ensemble of five *CNN*s, on the same *B-mode* and *M-mode* materials. (d) *Grad-CAM++* heatmaps localize the image regions driving the classification, letting experts judge whether the model attends to anatomically meaningful areas.

On the balanced *48*-case benchmark drawn from the independent hold-out set, the ensemble achieved ***100.0*% sensitivity (*95*% CI: *85.8–100.0*%)**, ***100.0*% specificity (*95*% CI: *85.8–100.0*%),** and ***100.0*% accuracy (*95*% CI: *92.6–100.0*%)**. With *24* positives the exact interval around a perfect point estimate still extends to *85.8*%, which is why intervals are presented throughout. **Performance was consistent** across tissue types (***100.0*% sensitivity on cadaveric and living-patient scans**) and on the difficult cases: the five cases the expert panel found hardest were living breath-hold mimics, and all five were classified correctly.

The *11* **experts** showed high but **variable sensitivity** (*B-mode 95.8*%, *95*% CI: *92.7–97.9*%; *M-mode 92.8*%, *95*% CI: *89.0–95.6*%), while **specificity proved harder** (*B-mode 81.0*%, *95*% CI: *75.8–85.6*%; *M-mode 73.5*%, *95*% CI: *67.7–78.7*%; individual range *45.8–100*%). In mixed-effects regression the sensitivity difference was not significant ($p = 0.095$ *M-mode*; $p = 0.134$ combined), but the *AI*'s specificity advantage was ($p = = 0.0247$ *M-mode*; $p = 0.0194$ combined). **No single expert achieved both perfect sensitivity and specificity**: six reached *100*% sensitivity with at best *87.5*% specificity; one reached *100*% specificity with *83.3*% sensitivity. Experts were significantly better in *B-mode* than *M-mode* for specificity ($p = = 0.00337$); the ensemble used *M-mode* exclusively and exceeded expert *B-mode* specificity—the information required is fully encoded in the time–motion representation. Inter-rater agreement was substantial for *B-mode* (*Fleiss'* $\kappa = 0.695$), moderate for *M-mode* ($\kappa = 0.577$). **With expert sensitivity near its ceiling, the clinical value lies in eliminating false positives, which carry the morbidity of needle decompression or chest-tube insertion.**

Explainability was rated, not asserted: five independent clinicians rated *84.2*% (*95*% CI: *79.0–88.2*%) of the *Grad-CAM++* **heatmaps clinically relevant**, with substantial

agreement (*Fleiss' κ = 0.750*); attention corresponded to the pleural line and the zone beneath it, where the seashore, barcode, and *T-line* signs (equivalent of lung pulse) are read.

4.2 Objective 2: feasibility and reliability of the BLUE-LUSS protocol

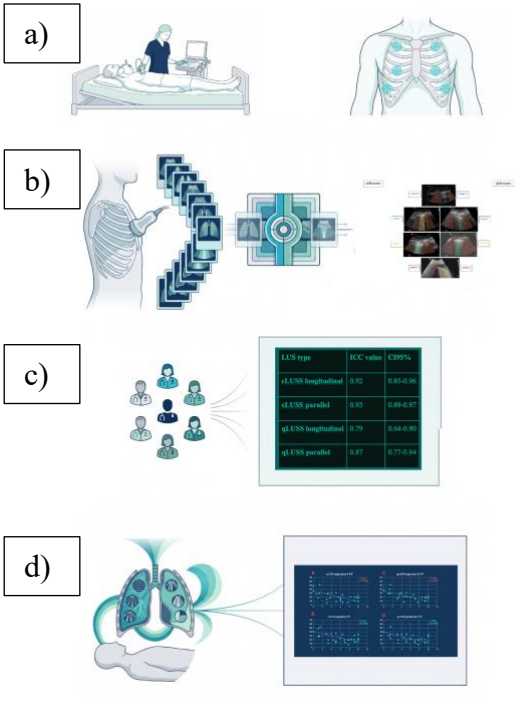

Figure 2: Workflow of Objective 2—feasibility and reliability of the BLUE-LUSS protocol in mechanically ventilated COVID-19 patients. (a) Single-operator bedside acquisition at the standardized *BLUE* points (upper, lower, and *PLAPS*, bilaterally), requiring no repositioning and no additional staff. (b) Each point is recorded in longitudinal and transverse orientations; the loops enter a standardized offline pipeline and are graded against a scoring atlas, separately for *cLUSS* and *qLUSS*. (c) Four blinded intensive care specialists score the full loop set: summarized *cLUSS* shows almost perfect, *qLUSS* good-to-excellent agreement. (d) Validation against gas exchange: all four score variants correlate inversely with the PaO_2/FiO_2 ratio, strongest for longitudinal *qLUSS*.

Twenty of *24* enrolled patients were analyzed (*400* loops, median age *60* years, *75*% male, *70*% unvaccinated, *50*% with ≥ *50*% lung involvement on *CT*, median *APACHE II 11.25*, median *CRP 163.2* mg/L). **Reliability of summarized scores was excellent:** *ICC 0.92* (*95*% CI: *0.85–0.96*) and *0.93* (*95*% CI: *0.89–0.97*) for longitudinal and transverse *cLUSS*; *0.79* (*95*% CI: *0.64–0.90*) and *0.87* (*95*% CI: *0.77–0.94*) for *qLUSS*. Individual loop scoring showed moderate agreement (κ = *0.56 cLUSS,* κ = *0.48 qLUSS*), consistent with published single-image variability—**aggregation stabilizes the measurement**, so **summarized scores suit serial monitoring**. Lung sliding detection agreed excellently (κ = = *0.92, 95*% CI: *0.78–1.00*).

All four **scoring variants correlated inversely with *PaO_2/FiO_2***: longitudinal *qLUSS* r = *-0.55* (p = *0.0119*), transverse *qLUSS* r = *-0.50* (p = *0.0235*), longitudinal *cLUSS* r = = *-0.48* (p = *0.0321*), transverse *cLUSS* r = *-0.46* (p = *0.0402*). **No correlation with C-reactive protein** (all p > *0.56*) and only weak-to-moderate correlations with white blood cell count—the score reflects aeration, not systemic inflammation. **The single-operator protocol proved feasible within routine *ICU* care under pandemic conditions**.

4.3 Objective 3: automated image processing and the bedside application

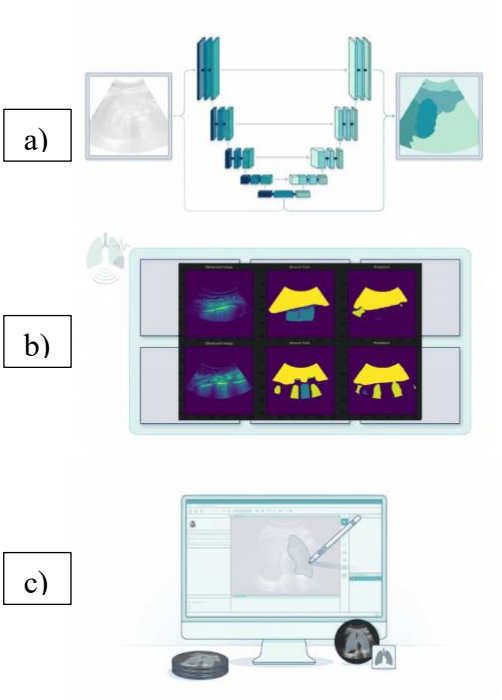

Figure 3: Workflow of Objective 3—automated semantic segmentation of lung ultrasound images as the basis of bedside image processing. (a) *U-Net* encoder–decoder: the *B-mode* input is downsampled along the contracting path while skip connections preserve spatial detail, yielding a pixel-wise multi-class map separating thoracic wall, ribs, rib shadows, pleural line, A-lines, B-lines, and consolidations. (b) Representative leave-one-out validation cases (original image, expert ground truth, model prediction), strong on prominent structures, weaker on early or subtle B-lines. (c) The *3D Slicer* annotation workstation: every frame manually delineated by an expert—the labor intensity of annotation, not model capacity, is the principal limit on dataset expansion.

The *U-Net* achieved *87.53*% **mean validation accuracy** under leave-one-out cross-validation at patient level—deliberately conservative: a random image-level split would have reported misleadingly higher accuracy. Performance tracked structural definition: excellent for thoracic wall and rib shadows, *70–80*% for subtle findings (early B-lines, pleural-line irregularity). Early overfitting indicated a data constraint, not an architectural one—pixel-level expert annotation is the rate-limiting step. **Lung sliding classification from *M-mode* exceeded 90%, establishing the route from static frames to temporal analysis.**

The work then progressed to a **functioning bedside application**. A stable Pi-local workflow ran live capture, *YOLO* localization, *M-mode* generation, ensemble classification, and optional *Grad-CAM++*; a native dashboard proved more trustworthy for validation than a browser client. Iterative optimization produced the first clearly usable Pi-local baseline in *March 2026*, and a hardware-accelerated branch integrated compiled *YOLO* and model ensemble *HEF* assets into a hybrid runtime. An automation variant ran a timed detection window, derived a confidence-weighted consensus scanline and executed the remaining steps automatically, with fail-safe interruption and manual fallback. A feedback layer stored clinician review together with session provenance, turning bedside sessions into traceable, analysis-ready records.

4.4 Objective 4: the validated cadaveric pneumothorax model

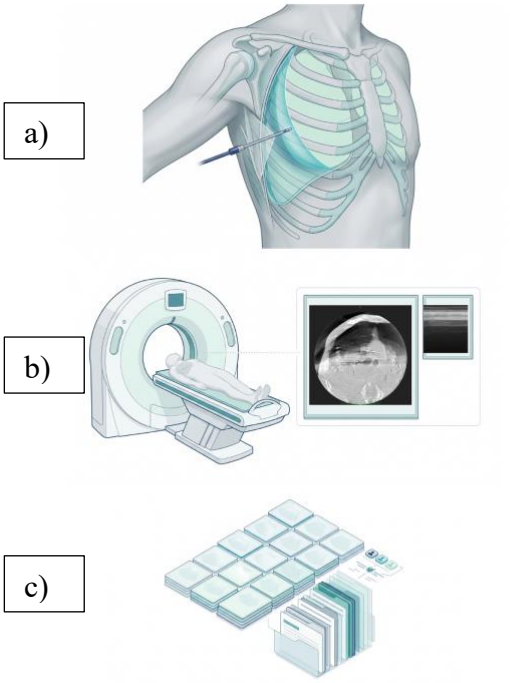

Figure 4: Workflow of Objective 4—development and validation of the human cadaveric pneumothorax model. (a) Minimally invasive induction: air insufflated through a small-bore cannula produces a controlled anterior air layer with the thoracic wall intact, keeping all *BLUE* points accessible—the key difference from more invasive models. (b) Verification: the sonographically determined borders are marked with metallic pins and confirmed by on-site *CT* (left), while the *M-mode* image documents the barcode sign at the same location (right)—an imaging-verified ground truth for each induced state. (c) Blinded expert evaluation of the *141* clips with a structured form (image quality, educational utility, *BLUE* profile identification, *LUSS* applicability); the clip library serves clinician training and *AI* ground truth alike.

Induction succeeded in all five cadavers of diverse body habitus, with the thorax intact and all *BLUE* points accessible. **The model reproduced the key signs**—absent lung sliding, A-profile dominance, barcode sign in *M-mode*—and the pathognomonic lung point was demonstrable in partial pneumothorax. From *141* **clips assessed** by two blinded intensivists, *92.9*% (*131/141*) **were rated acceptable for diagnostic purposes**; agreement was moderate for image quality ($\kappa = 0.524$), substantial for educational utility ($\kappa = 0.763$), moderate for *BLUE* profile identification ($\kappa = 0.627$) and scoring applicability ($\kappa = 0.631$). The stated limitation is absent circulation: no lung pulse, so the differential from other causes of absent sliding cannot be shown—itself an educational point. **Beyond training, the evolved model generated** the *826 CT*-verified *M-mode* images constituting the positive class for Objective 1—a data gap the field could not fill from clinical practice.

4.5 Objective 5: the integrated clinician-in-the-loop framework

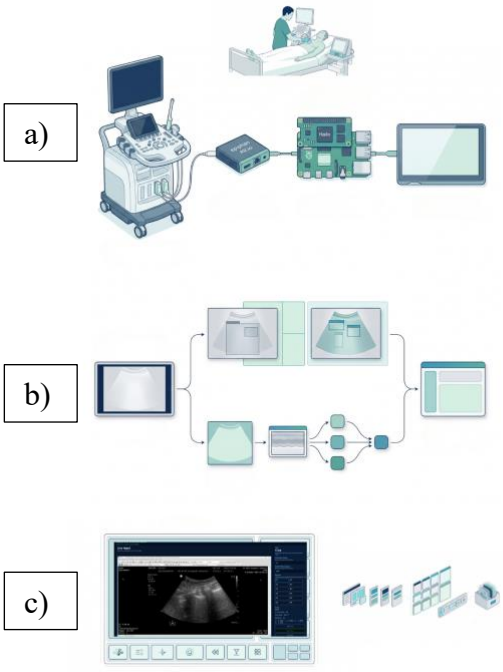

Figure 5: Workflow of Objective 5—the proposed translational framework realized as a deployable bedside decision-support system. (a) Hardware chain: the ultrasound video output passes through an *Epiphan AV.io* capture interface into a *Raspberry Pi 5* with *Hailo AI* accelerator and bedside display; the entire inference path is local—no cloud connection, no modification of the ultrasound device. (b) Edge pipeline: *YOLO*-supported localization identifies the pleural region and confirms the scanline, while the cropped region of interest is converted to *M-mode* and classified by the ensemble members, fused into a single soft-voting result on the clinician-facing dashboard. (c) The application interface (live input, identifiers, *BLUE*-region selection, crop definition) and the exported session record (region, *M-mode* image, ensemble output, *Grad-CAM++* status, clinician review)—the provenance layer that keeps the clinician in the loop.

The translational synthesis converged on a reproducible acquisition standard (depth at pleura + *3–8* cm, single focus at the pleura, artifact-targeted gain, enhancement features disabled, convex probes as default) **and four prerequisites of translational robustness: dataset diversity, expert-panel benchmarking, explainability on verifiable landmarks, and preserved clinician oversight.** The framework was realized as a deployable bedside architecture with a fully local inference path—no cloud, no modification of the ultrasound device—in which the clinician reviews the localization score, the ensemble label, and the explanation realism before the session is exported with full provenance. The result is not a conceptual proposal but a working decision-support pathway unifying acquisition discipline, edge deployment, automation, human review, and traceable export.

4.6 Integrated analysis and limitations

The studies are not independent: Objective 4 supplied the verified ground truth for Objective 1; the acquisition discipline of Objective 2 became the settings standard of Objective 5; the localization work of Objective 3 evolved into the runtime that closes the chain. By readiness: **three results are available for adoption now** (*BLUE-LUSS*, the cadaveric model, the vendor-independent acquisition protocol); **three are ready for prospective evaluation** (the ensemble as second reader—endpoint: unnecessary chest-tube insertion, not detection rate—, the bedside runtime, the explainability layer); **clinical deployment, regulatory conformity, and autonomous interpretation are explicitly not yet established.**

The principal limitation is the composition of the pneumothorax-positive class: predominantly *CT*-verified cadaveric preparations and only two living patients, so that exploitation of cadaver-versus-living differences cannot be excluded on statistical grounds alone. Two findings argue against it—correct

rejection of living breath-hold mimics, and *Grad-CAM++* attention localized to the pleural line. Further limitations: single-center design, finite expert panel, *M-mode* restriction, post-mortem physiology (no lung pulse), and segmentation dataset size.

5. Conclusions

This doctoral thesis delivers five contributions, each validated in its own study and each supplying a prerequisite for the next.

1. Verified ground truth. A minimally invasive, *CT*-verified human cadaveric pneumothorax model, accepted by independent experts (*92.9*% of clips acceptable; educational utility $\kappa = 0.763$), generating the labelled data—*826 CT*-verified images—the field did not have a priori.

2. A unified protocol. *BLUE-LUSS*, a single-operator semi-quantitative protocol for ventilated patients: excellent inter-rater reliability (ICC *0.92–0.93* for *cLUSS*), significant inverse correlation with oxygenation (r = *-0.55*), feasible within routine intensive care.

3. The signs, mapped. Multi-class semantic segmentation and live localization of the sonographic signs, validated on unseen patients (*87.53*% at patient level)—the route from retrospective algorithms to deployable point-of-care software.

4. Benchmarked and explained. An ensemble benchmarked against *11* expert clinicians—perfect sensitivity and specificity on the balanced benchmark, significant specificity superiority (p = *0.019*), *84.2*% of explanations rated clinically relevant—an auditable second reader that may reduce unnecessary chest-tube insertion.

5. Working at the bedside. A clinician-in-the-loop edge runtime with hardware acceleration, automation under explicit

fail-safes, structured review and provenance-aware export, publicly archived and citable (Zenodo DOI: 10.5281/zenodo.19541650).

No single result is decisive. Together, these support one claim: bedside reliability is not achieved by improving the algorithm alone—it requires acquisition, ground truth, algorithm, explanation, and review to be standardized as one chain. That is the sense in which the title should be read: not a device but a bedside pathway in which standardized acquisition, realistic training material, explainable inference, and clinician oversight operate together. The next step is prospective multicenter validation in living patients, which is the prerequisite for clinical deployment.

6. Bibliography of the candidate's publications

Publications related to the thesis:

Orosz G, Szabó RZ, Szabó M, Gyombolai P, Tóth JT, Ruttkay T, Ferenci T, Ungi T, Fichtinger G, Haidegger T. Explainable transfer learning ensemble AI model for lung ultrasound pneumothorax detection with expert benchmark. SCANDINAVIAN JOURNAL OF TRAUMA RESUSCITATION AND EMERGENCY MEDICINE (2026), in press. SJR: D1

Ádám N, Orosz G, Berczi M, Ruttkay T. Humán kadávermodell a légmell ágy melletti ultrahang-diagnosztikájához: Előtanulmány. ORVOSI HETILAP 164(46):1824–1830 (2023). SJR: Q4

Orosz G, Szabó RZ, Ungi T, Barr C, Yeung C, Fichtinger G, Gál J, Haidegger T. Lung Ultrasound Imaging and Image Processing with Artificial Intelligence Methods for Bedside Diagnostic Examinations. ACTA POLYTECHNICA HUNGARICA 20(8):69–87 (2023). SJR: Q2

Szabó M, Orosz G, Iványi ZD, Darvas K. A mellkas és a tüdő ultrahangvizsgálatának szerepe az általános érzéstelenítés során. ORVOSI HETILAP 164(22):864–870 (2023). SJR: Q4

Szabó RZ, Orosz G, Ungi T, Barr C, Yeung C, Incze R, Fichtinger G, Gál J, Haidegger T. Automation of lung ultrasound imaging and image processing for bedside diagnostic examinations. In: IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI 2023), Budapest, pp. 779–783 (2023).

Orosz G, Gyombolai P, Tóth JT, Szabó M. Reliability and clinical correlations of semi-quantitative lung ultrasound on BLUE points in COVID-19 mechanically ventilated patients: The 'BLUE-LUSS'—A feasibility clinical study. PLOS ONE 17(10):e0276213 (2022). SJR: Q1

Publications not related to the thesis:

Bódis F, Orosz G, Tóth JT, Szabó M, Élő LG, Gál J, Élő G. Percutaneous tracheostomy: Comparison of three different methods with respect to tracheal cartilage injury in cadavers—Randomized controlled study. PATHOLOGY AND ONCOLOGY RESEARCH 29:1610934 (2023). SJR: Q2

Bódis F, Orosz G, Szabó M, Molnár V, Tóth TJ, Élő LG, Tamás L, Élő G. Az antropometriai paraméterek szerepe a cadaveren végzett percutan tracheostomás módszerekben. ORVOSI HETILAP 164(16):630–635 (2023). SJR: Q4

∑IF: 13.4 (publications related to the thesis: 10.3; publications not related to the thesis: 3.1)