SEMMELWEIS EGYETEM

DOKTORI ISKOLA

Ph.D. értekezések

3533.

OROSZ GÁBOR

Nép- és közegészségtudományok

című program

Programvezető: Dr. Ács Nándor, egyetemi tanár

Témavezető: Dr. Haidegger Tamás, egyetemi tanár

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

Semmelweis University Doctoral College

Health Sciences Division

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 Complex Examination Committee: Tamás Ferenci, PhD, habil.

Enikő Kovács MD, PhD

Budapest

2026

List of abbreviations

POCUS – Point-of-Care Ultrasound
CT – Computed Tomography
ICU – Intensive Care Unit
ARDS – Acute Respiratory Distress Syndrome
BLUE – Bedside Lung Ultrasound in Emergency
PLAPS – Posterolateral Alveolar and/or Pleural Syndrome
LUSS – Lung Ultrasound Score
qLUSS – Quantitative LUSS
cLUSS – Coalescent LUSS
FALLS – Fluid Administration Limited by Lung Sonography
AI – Artificial Intelligence
CNN – Convolutional Neural network
XAI – Explainable AI
EHR – Electronic Health Record
DICOM – Digital Imaging and Communications in Medicine
Grad-CAM – Gradient-weighted Class Activation Map
CRP – C-Reactive Protein
PCT - Procalcitonin
WBC – White Blood Cell
APACHE-II - Acute Physiology and Chronic Health Evaluation
CURB-65 – Pneumonia Severity Score
PEEP – Positive End-Expiratory Pressure
GDPR – General Data Protection Regulation
THI – Tissue harmonic imaging
XRES – Speckle Noise Reduction
CrossXBeam – Spatial Compounding
SRI – Speckle Reduction
MI – Mechanical Index
TI – Tissue Thermal Index
SGD – Stochastic Gradient Descent
RMSProp – Root Mean Square Propagation
Raspberry Pi – Single-board Computer
HAILO AI HAT or abbrev. HAILO – High-performance Expansion Board (hardware attached on top)
ONNX – Open Neural Network Exchange
HEF – Hailo Execution Framework
ROI – Region Of Interest
ICC – Intraclass Correlation Coefficient
ROC – Receiver Operating Characteristic Curve
CI – Confidence Interval
CPU – Central Processing Unit
GPU – Graphics Processing Unit

1. Introduction

The title of this doctoral work refers to a lineage rather than to a device. 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 POCUS occupies the same position today: a portable, radiation-free extension of the bedside examination that allows the clinician to look at the chest in real time, at the moment the decision has to be made. The step examined in this thesis is the next one in that lineage. An ultrasound system coupled with an explainable artificial intelligence model — one that recognizes pleural motion, has been benchmarked against expert consensus, and can show the clinician which image region drove its output — is no longer only a probe; it functions as an „intelligent stethoscope". The decisive point, however, is not the instrument. Auscultation became clinically valuable because it was embedded in a disciplined routine, and an intelligence-assisted probe is valuable only under the same condition: standardized acquisition settings and scanning points, realistic training material, transparent inference, and a workflow in which the clinician retains review and final responsibility. This dissertation therefore treats the intelligent stethoscope not as a piece of hardware but as a complete bedside pathway, and its five objectives follow that pathway from protocol standardization and training-model development, through explainable machine learning model, to a deployable clinician-in-the-loop bedside system.

1.1 The evolution of lung ultrasound in critical care

The integration of ultrasound technology into critical care represents a paradigm shift in bedside diagnostics, with lung-focused Point-of-Care Ultrasonography (POCUS) standing out as a transformative innovation (1-11). Initially met with skepticism due to the perceived limitations of imaging air-filled organs, lung ultrasound gained traction in the early 2000s as evidence mounted regarding its ability to detect pathological patterns through artifact analysis (12). Pioneering work by clinicians such as Lichtenstein demonstrated that sonographic signs—like the absence of lung sliding for pneumothorax or the presence of B-lines for interstitial syndrome—could rival the diagnostic accuracy of traditional radiography and Computed Tomography (CT) in specific scenarios (13-18). This was particularly relevant in Intensive Care Units (ICUs), where rapid assessment of respiratory failure is critical, and transporting unstable patients for CT scans poses

significant risks. The advent of portable and handheld ultrasound devices further accelerated adoption, enabling real-time imaging at the bedside without ionizing radiation (19). During the COVID-19 pandemic, lung POCUS became indispensable for monitoring lung involvement in mechanically ventilated patients, reducing reliance on chest X-rays and minimizing healthcare workers' exposure (20-22). Studies have shown that lung ultrasound can detect COVID-19-associated pneumonia with significant sensitivity, facilitating early intervention and resource allocation (23-29). Beyond infectious diseases, its applications have expanded to include trauma, cardiogenic pulmonary edema, and Acute Respiratory Distress Syndrome (ARDS), where it guides fluid management and ventilator settings (10, 30-32). The evolution of lung POCUS reflects a broader trend toward minimally invasive, dynamic monitoring in critical care, underscoring its role in improving patient outcomes through timely, targeted therapies.

1.2 Standardized examination protocols: BLUE and quantitative scoring systems

Standardization is the cornerstone of reliable lung POCUS, and protocols like the Bedside Lung Ultrasound in Emergency (BLUE) protocol have been instrumental in systematizing examinations (2, 4, 5, 8, 33, 34). The BLUE protocol delineates specific points on the thorax—upper and lower BLUE points anteriorly, and the Posterolateral Alveolar and/or Pleural Syndrome (PLAPS) point posteriorly—allowing for consistent image acquisition across patients (3, 35). Each point is assessed for key artifacts:

- A-lines (horizontal reverberations) indicate normal aeration, while
- B-lines (vertical comet-tail artifacts) suggest interstitial fluid or consolidation,
- fractal/shred/tissue-like sign („shredded line" as a demarcation around non-translobar and translobar consolidations) indicates fully de-aerated regions.

The protocol further classifies findings into profiles (e.g., A-profile, B-profile, A/B-profile, C-profile, PLAPS-profile) that correlate with common causes of acute respiratory failure, such as Chronic Obstructive Pulmonary Disease (COPD) exacerbations, pneumonia, or pulmonary embolism. For completeness, the C-profile denotes an anterior consolidation pattern in which the normally aerated lung becomes tissue-like in appearance, often with hepatization-like echotexture: clinically, this pattern is most commonly associated with pneumonia or severe atelectatic states. This structured approach reduces diagnostic time and enhances accuracy, based on studies reporting that the BLUE protocol can identify the etiology of respiratory failure in over 90% of cases

when performed by trained operators. Building on this, quantitative scoring systems like the Lung Ultrasound Score (LUSS) provide objective measures of lung aeration. LUSS typically evaluates multiple regions (e.g., 12-zone protocol), assigning scores from 0 (normal aeration) to 3 (consolidation) based on artifact severity. Variants such as the coalescence-based LUSS (cLUSS) and quantitative LUSS (qLUSS) offer nuanced assessments, with qLUSS estimating the percentage of pleural involvement (36-38). These scores have been validated against CT scans and oxygenation parameters, showing strong correlations with PaO_2/FiO_2 ratios in ARDS patients (39-42). The clinical value of quantified LUS lies in its ability to transform a predominantly visual and operator-dependent examination into a reproducible and trendable monitoring tool. Numerical scoring improves inter-observer consistency, supports serial assessment of lung aeration loss or recovery, and provides an objective bridge towards clinical auditing and computer-assisted interpretation. Moreover, protocols like the Fluid Administration Limited by Lung Sonography (FALLS) adapt lung ultrasound for hemodynamic management. Such standardization not only supports diagnostic consistency but also enables multicenter research and tele-ultrasound applications, fostering collaboration and data sharing across institutions.

LUS is particularly valuable in anesthesiology, emergency care, and intensive care because it provides immediate, bedside, radiation-free and repeatable information at moments when respiratory deterioration, peri-intubation events, recruitment maneuvers, fluid therapy, or post-operative pulmonary complications require rapid decisions. In this setting, lung POCUS functions not merely as an imaging modality, but as a real-time extension of physical examination and a practical tool for immediate treatment guidance.

1.3 Clinical challenges and limitations in lung POCUS implementation

Despite its advantages, the implementation of lung POCUS faces multifaceted challenges that hinder its universal adoption. Operator dependency is a primary concern, as diagnostic accuracy is inextricably linked to the clinician's expertise and experience. Inter-observer variability studies reveal substantial discrepancies even among experts; for instance, interpretations of B-line patterns can vary by up to 20% among practitioners, leading to inconsistent management decisions (43). This variability stems from the subjective nature of artifact interpretation, where factors like transducer pressure, angle, and machine settings (e.g., gain, depth) influence image quality and perception. Novice

operators often struggle with distinguishing true pathologies from other artifacts, such as misidentifying Z-lines (short, non-pathological reverberations) for B-lines, which can result in false positives. While basic competency is attainable after roughly 25 supervised examinations, recommended training thresholds vary several-fold across societies, and no unified certification standard exists across specialties, yet ongoing practice is essential to maintain skills - a challenge in resource-limited settings, where training opportunities are scarce (44). Additionally, patient-related factors complicate examinations: obesity, subcutaneous emphysema, or dressings can obscure acoustic windows, while anatomical variations affect the reproducibility of standardized points. In critical care environments, time constraints and high cognitive loads exacerbate these issues, increasing the risk of errors during rapid assessments. Furthermore, the lack of universal guidelines for machine calibration and protocol adherence leads to institutional disparities in data quality (45). For example, differences in depth settings between devices can alter the apparent density of B-lines, affecting score calculations. These limitations underscore the need for decision-support tools and enhanced training methodologies to reduce subjectivity and improve reliability across diverse clinical scenarios.

1.4 The promise of artificial intelligence in lung POCUS

Artificial Intelligence (AI) holds immense potential to overcome the limitations of human-dependent lung POCUS by automating image analysis and enhancing diagnostic precision. Deep learning models, particularly Convolutional Neural Networks (CNNs), excel at recognizing complex patterns in ultrasound images (46). Transfer learning approaches, where pre-trained networks are fine-tuned on lung ultrasound data, accelerate model development and improve generalizability (47-49). Ensemble methods, which aggregate predictions from multiple architectures, further boost performance by reducing overfitting and leveraging complementary features (50-52). Explainable AI techniques become integral parts of trustworthy systems (53-55). Beyond diagnosis, AI applications include automated scoring systems that quantify lung aeration in real-time, providing objective measures for monitoring disease progression. For example, AI algorithms can track B-line dynamics over time, correlating with fluid status in heart failure patients(56). Integration with Electronic Health Records (EHRs) allows for predictive analytics, such as forecasting the need for mechanical ventilation based on serial ultrasound findings. However, challenges remain, including the need for diverse, annotated datasets that

encompass rare pathologies and varying patient demographics. Data augmentation techniques—such as rotation, scaling, and synthetic image generation—help mitigate dataset biases, while federated learning approaches enable model training across institutions without data sharing (57). As AI tools evolve, their deployment on edge devices (e.g., tablets, handheld ultrasound systems) promises point-of-care decision support, ultimately standardizing interpretations and reducing the diagnostic burden on clinicians.

1.5 Integrating bioengineering for translational impact

Translational bioengineering bridges the gap between technological innovation and clinical practice by optimizing the entire lung POCUS ecosystem. A key focus is dataset curation, which involves assembling real-world multi-source images—if available. Cadaveric simulations, for instance, provide controlled environments for generating pathological signs like pneumothorax, enabling the collection of ground-truth data for AI training without patient risk. Standardizing image acquisition parameters is another critical area; bioengineers collaborate with clinicians to define optimal settings that maximize artifact clarity and reproducibility. This standardization supports the development of automated quality control algorithms that flag suboptimal images in real-time, prompting operators to adjust acquisition techniques. On the hardware front, portable devices with embedded AI capabilities are being engineered for seamless integration into clinical workflows, featuring user-friendly interfaces that guide novice users through protocol steps. Software innovations include DICOM-compliant platforms for data storage and analysis, as well as cloud-based systems for tele-ultrasound and remote expertise sharing. Additionally, bioengineering efforts address ethical and practical concerns, such as data privacy through encryption and anonymization, and scalability through low-cost solutions for resource-limited settings. By harmonizing engineering principles with clinical needs, translational bioengineering aims to create end-to-end solutions that enhance the accessibility, accuracy, and impact of lung POCUS worldwide.

1.6 Identifying research gaps and thesis objectives

While significant progress has been made, several research gaps impede the full integration of AI and bioengineering into lung POCUS. First, AI models often lack

generalizability due to training on homogenous datasets that underrepresent diverse populations, pathologies, and imaging conditions. For example, models trained on data from single centers may fail in multicultural settings or with uncommon conditions like pleural effusions in obese patients. Second, validation against human performance is frequently inadequate, with few studies employing large, multidisciplinary expert panels for benchmarking. This limits the clinical translation of AI tools, as real-world variability is not fully captured. Third, the practical deployment of AI in bedside settings faces hurdles related to workflow integration, real-time processing speeds, and user acceptance. Clinicians may resist AI assistance due to trust issues or increased cognitive load, highlighting the need for intuitive interfaces and robust validation. Fourth, the development of realistic training models for education and validation remains nascent; existing phantoms and simulators often lack the fidelity to replicate complex artifacts or dynamic physiological processes. Finally, ethical considerations—such as data security, algorithm transparency, and equity in access—require comprehensive frameworks to ensure responsible innovation. This thesis aims to address these gaps through a multidisciplinary approach that combines clinical expertise with engineering solutions. Specific research objectives include:

- Developing an explainable, ensemble AI model for pneumothorax detection trained on a diverse dataset including clinical, cadaveric, and volunteer images, and rigorously validating it against a panel of 11 experienced clinicians to establish superiority in sensitivity and specificity.
- Assessing the feasibility and reliability of the BLUE-LUSS protocol at mechanically ventilated COVID-19 patients, examining inter-observer agreement and correlations with oxygenation parameters to refine simplified scoring systems for critical care.
- Creating and validating a human cadaveric model for pneumothorax simulation, evaluating its utility for both clinician training and AI data augmentation through expert-rated image quality and diagnostic accuracy.
- Proposing standardized imaging protocols and bioengineering frameworks for integrating AI tools into POCUS workflows, focusing on real-time decision support, data standardization, and ethical deployment.

By addressing these objectives, this thesis seeks to advance lung POCUS toward a future where it is more standardized, accurate, and accessible, ultimately improving patient outcomes and resource efficiency in critical care.

2. Objectives

This thesis aims to systematically address the critical challenges in lung-focused POCUS by integrating evidence-based bedside protocols with innovative translational bioengineering approaches. The research objectives have been developed through a comprehensive analysis of clinical needs and technological opportunities, with each objective building upon the findings from previous investigations. The multi-faceted approach encompasses AI development, clinical protocol validation, educational model creation, and implementation framework establishment. Each objective is coupled with a specific, testable hypothesis derived from empirical observations and preliminary data, ensuring a rigorous scientific foundation for the research program. The following sub-sections provide detailed descriptions of the research aims, methodological approaches, and expected outcomes based on the comprehensive work conducted across multiple studies.

2.1 Objective 1: Develop and validate an explainable AI ensemble model for pneumothorax detection (58)

Background and rationale

Pneumothorax represents one of the most critical diagnostic emergencies in intensive care settings, with tension pneumothorax potentially leading to cardiovascular collapse within minutes if not promptly identified and treated (59). Traditional diagnostic approaches, including supine chest X-rays, demonstrate suboptimal sensitivity, particularly in critically ill patients where clinical examination may be unreliable. Lung ultrasound has emerged as a rapid, radiation-free alternative with diagnostic accuracy approaching that of CT, yet its implementation is severely limited by inter-observer variability and dependence on operator expertise. The fundamental challenge lies in the subjective interpretation of dynamic sonographic signs, including the absence of lung sliding (60), the presence of the barcode sign in M-mode (61), and the identification of the pathognomonic lung point (62). Previous attempts at automated detection have been hampered by insufficiently diverse training datasets, inadequate validation against expert performance, and lack of clinical interpretability in AI models (63). This objective addresses these limitations through the development of a comprehensive AI system that leverages ensemble learning, transfer learning, and explainable AI techniques to create a

robust diagnostic tool that can function as a reliable second reader in critical care environments.

Detailed objective description

The primary aim is to design, implement, and rigorously validate a soft-voting ensemble of five convolutional neural network architectures—ResNet, Inception v3, Xception, Inception-ResNet-v2, and VGG16—specifically optimized for pneumothorax classification using M-mode lung ultrasound images. The model development incorporated transfer learning from ImageNet pre-trained weights, followed by comprehensive fine-tuning on a diverse dataset of 1,856 M-mode images derived from multiple sources:

(i) critically ill patients with various respiratory pathologies from a university hospital ICU during and after the COVID-19 pandemic,

(ii) healthy volunteers performing breath-holding maneuvers to simulate challenging diagnostic scenarios, and

(iii) tailored human cadaver models with artificially induced pneumothorax under controlled conditions.

The dataset encompassed the full spectrum of relevant pleural motion patterns, including normal lung sliding (seashore sign), subtle lung pulse (T-lines), absent lung sliding (barcode sign), and the definitive lung point. A critical innovation involves the implementation of Gradient-weighted Class Activation Mapping (Grad-CAM++) for model explainability, generating heatmaps that highlight the image regions most influential in the classification decision. The validation protocol includes five-fold cross-validation on the training set and final evaluation on an independent hold-out test set comprising 371 images. Most importantly, the model's performance was benchmarked against a panel of 11 experienced clinicians specializing in point-of-care ultrasound, each with over ten years of LUS experience, using a balanced set of 48 cases not included in training. Statistical analysis encompassed sensitivity, specificity, accuracy calculations with exact Clopper-Pearson confidence intervals, and mixed-effects logistic regression to account for correlations within subjects and imaging modes(64-66).

Comprehensive hypothesis

The ensemble AI model achieved diagnostic performance superior to individual human experts, with point estimates of 100% sensitivity and 100% specificity on the balanced

test set, significantly reducing false-positive diagnoses compared to clinicians, particularly in challenging scenarios such as subtle pleural motions and breath-holding artifacts. The model demonstrated consistent performance across different ultrasound modes, maintaining perfect sensitivity regardless of whether experts prefer B-mode or M-mode interpretation, while significantly outperforming human specificity in M-mode imaging ($p<0.001$). The explainability component showed substantial inter-rater agreement among experts (Fleiss' $\kappa > 0.7$) regarding the clinical relevance of Grad-CAM++ heatmaps, with over 80% of highlighted regions corresponding to anatomically meaningful areas crucial for pneumothorax diagnosis. Furthermore, the model exhibited excellent generalizability, performing equally well on both cadaveric simulations and real patient scans, validating the use of controlled cadaver models for AI training and evaluation. The AI system effectively functioned as an "idealized expert committee", achieving consensus-level performance that no individual clinician can consistently maintain, thereby standardizing diagnostic accuracy and reducing operator-dependent variability in critical care settings.

2.2 Objective 2: Assess the feasibility and reliability of the BLUE-LUSS protocol in mechanically ventilated COVID-19 patients (67)

Background and rationale

The COVID-19 pandemic created unprecedented challenges in intensive care units worldwide, highlighting the urgent need for efficient, reproducible monitoring tools that could provide rapid assessment of lung pathology while minimizing healthcare worker exposure and resource utilization (68-70). Traditional comprehensive lung ultrasound protocols, such as the 12- or 14-region examinations, proved cumbersome in the context of mechanically ventilated patients, particularly those in prone position, requiring multiple staff members and significant time investment. The BLUE (Bedside Lung Ultrasound in Emergency) protocol offered a simplified approach with defined examination points but lacked quantitative assessment capabilities essential for serial monitoring. Simultaneously, existing quantitative scoring systems like the Lung Ultrasound Score (LUSS) required extensive scanning that was often impractical in overwhelmed ICU environments. This objective addresses the critical need for a balanced approach that combines the efficiency of the BLUE protocol with the quantitative power

of LUSS systems, creating a feasible tool for daily assessment of mechanically ventilated COVID-19 patients while maintaining reliability and clinical correlation.

Detailed objective description

This objective involves the prospective validation of the novel BLUE-LUSS protocol—an integration of traditional BLUE point examination with semi-quantitative scoring systems including coalescence-based LUSS (cLUSS) and quantitative LUSS (qLUSS)—in a cohort of 24 consecutive SARS-CoV-2 RT-PCR positive, mechanically ventilated critically ill patients. The study design incorporates standardized ultrasound imaging at the three conventional BLUE points per hemithorax, with simultaneous acquisition of both longitudinal and transverse scans to compare their relative reliability and correlation with the clinical parameters. A comprehensive dataset of 400 ultrasound loops was recorded using strictly standardized machine settings (depth: pleural line + 3-8 cm; single focus at pleural line; mechanical index <0.7; tissue harmonic imaging disabled) across two ultrasound systems (Philips CX50 and GE Venue GO) with convex transducers. Four independent intensive care specialists, each with over seven years of LUS experience, performed blinded offline analysis using standardized scoring systems: cLUSS (0: A-lines/maximum two B-lines; 1: ≥3 well-spaced B-lines; 2: coalescent B-lines; 3: tissue-like pattern) and qLUSS (based on percentage of involved pleura).

The primary outcomes include inter-observer reliability assessed through intraclass correlation coefficients for summarized scores and Cohen's κ for individual loop scoring, with secondary outcomes focusing on correlations between BLUE-LUSS values and clinical parameters including PaO_2/FiO_2 ratio, inflammatory biomarkers (CRP, PCT, WBC), and predictive scores (APACHE II, CURB-65) (71, 72). The study is powered to detect intraclass correlation coefficients of 0.8 ± 0.15 with 80% power at α=0.05, requiring 20 patients, with 24 enrolled to account for potential exclusions.

Comprehensive hypothesis

The BLUE-LUSS protocol demonstrated excellent inter-rater reliability for summarized scores, with intraclass correlation coefficients exceeding 0.9 for cLUSS (both longitudinal and transverse) and 0.8–0.9 for qLUSS calculations, while individual loop scoring showed moderate agreement (κ 0.5-0.6) consistent with established LUS variability (73). The protocol will show significant inverse correlations with PaO_2/FiO_2 ratios across all scoring methods, with longitudinal qLUSS demonstrating the strongest

correlation (r = -0.55, p = 0.012), supporting its superiority at oxygenation assessment. The single-operator methodology proved feasible without requiring patient repositioning or additional staffing, with examination times compatible with routine ICU workflow. Interestingly, no significant correlations emerged between BLUE-LUSS scores and inflammatory biomarkers (CRP, PCT), reinforcing that lung ultrasound primarily reflects structural rather than systemic inflammatory changes. The combination of BLUE points with LUSS scoring maintained the diagnostic accuracy of more extensive protocols while reducing examination time and complexity, making it particularly valuable during pandemic conditions with limited resources and high infection risk. The protocol successfully bridged the gap between qualitative assessment and quantitative monitoring, providing a standardized approach for serial evaluation of lung aeration at mechanically ventilated patients with heterogeneous pulmonary pathology.

2.3 Objective 3: Automate lung ultrasound image processing using artificial intelligence for bedside diagnostics (74, 75)

Background and rationale

The interpretation of lung ultrasound images remains heavily dependent on clinician expertise, creating significant barriers to consistent implementation, particularly in emergency settings and resource-limited environments. The learning curve for proficient LUS interpretation is substantial, requiring recognition of complex artifact patterns and dynamic signs that may be subtle or context-dependent. Previous attempts at automation have typically focused on isolated aspects of lung ultrasound analysis, such as B-line counting or consolidation detection, without providing comprehensive decision support aligned with established clinical protocols like the BLUE protocol. Furthermore, most existing systems lack integration with clinical workflows, and fail to provide the explainability necessary for clinician trust and adoption. This objective addresses these limitations by developing an end-to-end AI system that progresses from low-level image processing to high-level diagnostic decision support, ultimately creating a foundation for automated bedside lung ultrasound interpretation.

Detailed objective description

This objective encompasses the implementation and validation of a U-Net-based deep neural network for multi-class semantic segmentation of lung ultrasound images, targeting the identification of primary signs essential for the BLUE protocol. The model

architecture incorporates zero-padding in convolutional layers to preserve input dimensions and employs a standard encoder-decoder structure with skip connections to maintain spatial information crucial for accurate segmentation. The training utilizes a comprehensively annotated dataset of 1,097 manually segmented ultrasound frames derived from 22 RT-PCR confirmed SARS-CoV-2 positive patients and 18 non-COVID patients, with expert annotations performed using 3D Slicer software (76). Segmentation classes include critical anatomical and pathological elements: thoracic wall, ribs, rib shadows, B-lines, septal rockets, and abnormal pleural line characteristics. The preprocessing pipeline involves standardization to 256×256 pixels quadratic format through resizing with linear interpolation for images and nearest-neighbor interpolation for masks, followed by masking to remove extraneous machine-generated information. Model training employs leave-one-out cross-validation based on patient-level splitting to ensure generalizability to new patients, with evaluation metrics including validation accuracy and sparse categorical cross-entropy loss. Beyond segmentation, the objective includes the development of classification networks for detecting secondary signs requiring temporal analysis, such as lung sliding, through analysis of B-mode video sequences transformed into M-mode representations. The research also explores the transition from research implementation to clinical deployment, including integration possibilities with portable ultrasound devices and real-time processing requirements for bedside use.

Comprehensive hypothesis

The U-Net segmentation model achieved mean validation accuracy exceeding 85% across cross-validation folds, with particularly strong performance for prominent anatomical features like thoracic wall and rib shadows and moderate performance for more subtle pathological findings like early B-lines and pleural line abnormalities. The leave-one-out validation demonstrated robust generalizability to new patient data, though performance was be limited by current dataset size, leading to early overfitting indications within initial training epochs. The semantic segmentation outputs provided reliable foundation for higher-level diagnostic tasks, with segmented features enabling automated calculation of LUSS scores and identification of BLUE protocol profiles. The transition from static frame analysis to dynamic sign detection showed promising results, with classification networks achieving >90% accuracy in distinguishing presence versus

absence of lung sliding from M-mode representations. However, the research identified significant challenges in scaling annotation efforts due to the labor-intensive nature of manual segmentation, suggesting future directions including alternative approaches like object detection with bounding boxes to accelerate dataset expansion. The integrated system demonstrated potential for real-time bedside decision support, though clinical deployment required additional validation and interface optimization to ensure seamless integration into clinical workflows without disrupting established examination patterns.

2.4 Objective 4: Develop and validate a human cadaveric model for pneumothorax simulation in ultrasound training (77, 78)

Background and rationale

Effective training in lung POCUS requires exposure to diverse pathological findings, yet opportunities for hands-on experience with critical conditions like pneumothorax are limited in clinical settings due to their emergent nature and potential patient harm. Existing training models include commercial phantoms that often lack anatomical fidelity and animal models that raise ethical concerns and differ significantly from human thoracic anatomy (79). Human cadaveric models offer a promising alternative by providing realistic tissue properties and anatomical relationships, but previous implementations have typically involved highly invasive approaches that compromise thoracic integrity and limit ultrasound examination feasibility. This objective addresses these limitations through the development of a minimally invasive cadaveric model that preserves thoracic anatomy while reliably reproducing sonographic signs of pneumothorax, serving dual purposes of clinician education and AI training data augmentation.

Detailed objective description

This objective involves the creation and validation of a human cadaveric pneumothorax model using fresh, unembalmed cadavers (n = 5) with intact thoracic anatomy and no traumatic injuries or prior thoracic surgeries. The model employs a minimally invasive approach involving endotracheal intubation with mechanical ventilation simulation using a resuscitation bag with PEEP valve, followed by precise insertion of a 3 mm internal diameter cannula into the pleural space at the “safety triangle” (fifth intercostal space, mid-axillary line) using blunt dissection to minimize tissue damage. Pneumothorax is induced through controlled air insufflation using a membrane pump capable of generating

pressures up to 140 cm H_2O, with the cannula temporarily clamped to maintain intrapleural air. Ultrasound examination follows the BLUE protocol points before and after pneumothorax induction, recording 10-second clips at each location using a GE LOGIQ V2 system with convex transducer (4C-RS, 2.0-5.5 MHz) and standardized settings. The generated image library (141 clips total) undergoes expert evaluation by two independent intensive care specialists with over five years of POCUS experience, assessing image quality, educational utility, clinical relevance for profile identification, and suitability for semi-quantitative scoring using a structured scoring system. Statistical analysis includes Cohen's κ with Fleiss-Cuzick extension for inter-rater agreement, with the study powered to detect $\kappa > 0.6$ with 80% power at $\alpha = 0.05$. The model also enables systematic variation of pneumothorax size through controlled air volume adjustment, allowing demonstration of progressive sonographic changes from subtle signs to tension pneumothorax.

Comprehensive hypothesis

The cadaveric model generated ultrasound images of moderate to excellent quality, with expert ratings showing substantial agreement (κ 0.5-0.7) across evaluation categories including image quality parameters, educational utility, and clinical relevance for BLUE profile identification. The model reliably reproduced key sonographic signs of pneumothorax, including absence of lung sliding, A-profile dominance, and barcode sign in M-mode, while the lung point was demonstrable in partial pneumothorax conditions (80). The minimally invasive approach preserved thoracic integrity sufficient for comprehensive ultrasound examination at all BLUE points, unlike more invasive models that compromise anterior examination sites. Experts rated the model as highly suitable for both novice training and expert refinement, particularly for recognizing challenging differentiations such as between true pneumothorax and conditions causing absent lung sliding like mainstem intubation or pleurodesis. The model additionally proved valuable for AI training data diversification, providing ground-truth pathological examples that address gaps in clinical data availability, particularly for progressive pneumothorax stages and controlled variations. However, limitations included the inability to replicate cardiogenic oscillations (lung pulse) due to absent circulation, creating an important educational point regarding differential diagnosis in clinical practice. The cost-effectiveness and reusability of the model supported scalable implementation in

simulation-based education programs, potentially reducing the time required for competency development in lung POCUS.

2.5 Objective 5: Propose comprehensive frameworks for integrating AI and bioengineering into lung POCUS workflows (58, 74, 75)

Background and rationale

The successful translation of technological innovations from research to clinical practice requires careful consideration of implementation frameworks that address workflow integration, data standardization, validation protocols, and ethical considerations. Disparities in image acquisition protocols, machine settings, and examination techniques currently limit the comparability of lung ultrasound studies across institutions and hinder the development of robust AI systems. Furthermore, the deployment of AI assistance in clinical settings introduces new challenges including user interface design, real-time processing requirements, result interpretation, and maintaining clinician autonomy. This objective addresses these comprehensive implementation challenges by developing evidence-based frameworks that standardize the entire lung POCUS pipeline from image acquisition to AI-assisted interpretation, ensuring consistent performance across diverse healthcare environments while maintaining ethical standards and clinical relevance.

Detailed objective description

This objective involves the development of comprehensive frameworks based on systematic analysis of image quality determinants, workflow patterns, and implementation barriers identified across previous studies. The image acquisition component establishes evidence-based protocols for machine settings including depth (pleural line + 3-8 cm), focus (single focus at pleural line with multi-focus disabled), gain optimization for primary artifacts, and specific disabling of image processing features (tissue harmonic imaging, speckle reduction, spatial compounding) that may artifactually alter diagnostic signs. The framework includes transducer selection guidelines based on patient characteristics and diagnostic questions, with convex arrays (1–5 MHz) recommended for most clinical scenarios and linear arrays reserved for detailed pleural assessment. For AI integration, the framework outlines validation requirements including dataset diversity metrics (covering patient demographics, pathology spectrum, imaging conditions), benchmarking against multi-expert panels, and clinical utility assessments measuring impact on diagnostic time, accuracy, and user confidence.

Comprehensive hypothesis

The implementation of standardized acquisition protocols significantly improved image consistency across operators and devices, with expert ratings showing improved interpretability and reduced variability in artifact characterization compared to non-standardized approaches. The integration framework demonstrated that user-centered interface design significantly impacts AI adoption, with simplified presentation of AI conclusions plus optional detailed explanation (via Grad-CAM++ heatmaps) achieving higher clinician acceptance compared to either simple outputs alone or overwhelming raw data presentation. The comprehensive framework reduced implementation barriers for AI-assisted lung POCUS, particularly in resource-limited settings where expert supervision is scarce, potentially expanding access to high-quality ultrasound interpretation while maintaining safety standards through appropriate human oversight and continuous performance monitoring.

The overall thesis structure, linked datasets, principal methods, and main outputs are summarized in Appendix Table 1.

3. Materials and methods

This chapter provides an exhaustive description of the integrated methodologies employed across the multidisciplinary studies constituting this thesis. The comprehensive approach spans clinical data collection, ultrasound imaging standardization, advanced AI development, and rigorous statistical analysis. Each methodological component was designed to ensure reproducibility, validity, and clinical relevance while addressing the complex challenges of lung-focused point-of-care ultrasonography in critical care settings.

3.1 Study populations and data collection

3.1.1 Clinical patient cohorts

Multiple distinct patient cohorts were strategically selected and utilized across the research program to ensure comprehensive representation of clinical scenarios and pathological conditions, as described below. For the primary AI model development and validation, an extensive dataset was prospectively collected from critically ill patients admitted to the intensive care unit of Semmelweis University during and after the extended COVID-19 pandemic period (2020-2024). This temporally diverse collection ensured representation across different pandemic waves and viral variants, capturing the evolving clinical presentations and management strategies. The inclusion criteria encompassed adults (>18 years) with various respiratory pathologies, creating a heterogeneous dataset that included acute respiratory failure due to COVID-19 pneumonia, bacterial pneumonia, ARDS, COPD exacerbations, and postoperative respiratory complications. This diversity was crucial for developing robust AI models capable of generalizing across multiple clinical scenarios rather than being optimized for a single disease entity.

The final curated dataset comprised 1,856 high-quality M-mode images derived from 114 unique subjects, carefully balanced across multiple categories. Pneumothorax-positive cases included 23 cadavers with artificially induced pneumothorax under controlled conditions and 2 live patients with radiologically confirmed spontaneous pneumothorax. Pneumothorax-negative cases included 79 living patients with various respiratory pathologies but definitively excluded pneumothorax, and 10 healthy volunteers providing normal baseline data. The cadaveric subgroup underwent a sophisticated preparation

protocol involving sequential air insufflation to simulate varying pneumothorax sizes, with each volume condition treated as repeated measurements of the same subject rather than independent cases. For living patients, pneumothorax absence was confirmed through comprehensive reference standards including CT scanning, serial chest X-rays, and consensus review by at least two independent intensivists with >10 years of experience. The data partitioning strategy employed strict subject-level splitting to prevent data leakage, with all clips from a given subject (patient, volunteer, or cadaver) assigned exclusively to either training or testing partitions, ensuring that the model evaluation truly reflected performance on previously unseen individuals.

For the BLUE-LUSS feasibility study specifically targeting COVID-19 patients, a prospective, observational design was implemented involving 24 consecutive SARS-CoV-2 RT-PCR positive, mechanically ventilated critically ill patients admitted between December 2021 and April 2022. This period strategically captured the transition between delta and omicron variants' dominance in Hungary, allowing observation of potential sonographic pattern variations between variants. Exclusion criteria were meticulously defined to include any living will prohibiting ICU care or declared treatment restrictions, ensuring ethical consistency. Comprehensive clinical parameters were systematically recorded at the time of ultrasound examination, including detailed anthropometric characteristics, validated predictive scores (APACHE II for illness severity, CURB-65 for pneumonia severity), serial laboratory findings (WBC count, CRP, PCT), dynamic oxygenation parameters (PaO_2/FiO_2 ratio), and detailed vaccination status with specific attention to timing relative to symptom onset. This multidimensional clinical characterization enabled sophisticated correlation analyses between sonographic findings and clinical parameters.

3.1.2 Cadaveric models

Two separate cadaveric cohorts were used in this work and should not be conflated. The first, comprising five fresh, unembalmed cadavers, served the preliminary feasibility study that established and validated the induction protocol. The second, comprising 23 prepared cadavers, provided the CT-verified pneumothorax-positive cases for the AI study; the cadaveric subjects referred to in Section 3.1.1 (n = 23) and counted among the 114 unique subjects belong to this second cohort. Within each cadaver, sequential air volumes were treated as repeated measurements of the same subject, and partitioning was

performed strictly at the cadaver level. The development of human cadaveric models represented a significant bioengineering innovation within this research program. Five fresh, unembalmed cadavers (3 females aged 79-88 years, 2 males aged 70-74 years) with intact thoracic anatomy and no traumatic injuries or prior thoracic surgeries were meticulously selected. Cadavers were donated specifically for educational and research purposes to the Department of Anatomy, Histology and Embryology at Semmelweis University through established body donation programs with full ethical compliance. The model development involved a sophisticated multi-step protocol beginning with proper positioning simulating supine ICU patients, followed by endotracheal intubation using appropriate-sized Macintosh blades and endotracheal tubes secured with standard fixation. Mechanical ventilation simulation employed a resuscitation bag with integrated PEEP valve set at 5 cmH_2O, delivering tidal volumes of 6-7 mL/kg at 12 breaths per minute to approximate protective lung ventilation strategies used in critical care.

The pneumothorax induction methodology represented a key innovation through its minimally invasive approach. A 3 mm internal diameter cannula was precisely inserted into the pleural space at the anatomically defined “safety triangle” (fifth intercostal space, mid-axillary line) using blunt dissection technique to minimize tissue damage and preserve thoracic integrity. This approach contrasted favorably with previously published methods that involved extensive surgical access or bronchial ligation, which compromised anatomical relationships and limited ultrasound examination feasibility. Pneumothorax was induced through controlled air insufflation using a Sera Precision Air 275 R plus (sera GmbH, Heinsberg, Germany) membrane pump capable of generating pressures up to 140cm H2O, with the cannula temporarily clamped to maintain intrapleural air. This system allowed precise titration of pneumothorax size and reproducible simulation of different clinical scenarios from small apical collections to tension physiology. The model’s versatility enabled demonstration of progressive sonographic changes and even allowed for drainage simulation by unclamping the cannula, illustrating lung re-expansion dynamics—a unique educational feature not available in commercial simulators.

To establish a robust foundation for the subsequent AI analysis, an evolved model was employed in the AI study on 23 subjects. This study utilized mechanically ventilated cadavers (Newport HT70 machine, Medtronic plc, Galway, IE) to simulate a dynamic

physiological environment. Following the artificial induction of a pneumothorax with three different amounts of air (e.g.: 200, 400 and 600 mL respectively), the pleural borders and its precise extent were first identified using ultrasound. These margins were demarcated on the cadaver's skin with permanent ink and subsequently marked with small metallic pins to ensure accurate correlation during computed CT imaging. Leveraging the unique resource of an on-site CT scanner within the Human Anatomy Department, cadavers were immediately scanned to validate that the externally marked borders faithfully represented the pneumothorax area (FIDEX GT, Animage LLC, Pleasanton, CA). All CT scans were then reviewed by an independent radiologist, who was blinded to the study's aims, to confirm both the presence and the correct delineation of the pneumothorax in relation to the metallic markers. Only after this confirmation were ultrasound images of the verified pneumothorax regions systematically acquired. Throughout the imaging process, the cadavers were maintained on standardized mechanical ventilation with volume-controlled settings, a tidal volume of 5 mL/kg of body weight, and a PEEP of 5 cmH_2O.

3.1.3 Healthy volunteers

Healthy volunteer recruitment targeted physicians with understanding of the research's requirements and ability to perform specific maneuvers. Ten volunteers contributed normal lung ultrasound recordings, with particular emphasis on generating challenging diagnostic scenarios through controlled maneuvers. Breath-holding at different lung volumes simulated absent lung sliding while maintaining lung-pulse, creating the classic "T-line" pattern that can be misinterpreted as pneumothorax by inexperienced operators. Varied respiratory patterns including rapid shallow breathing and occasional deep sighs generated a spectrum of normal lung sliding appearances. All volunteers provided comprehensive written informed consent specifically addressing ultrasound examination, anonymized data collection, secure storage, and subsequent analysis for both educational and research purposes, including AI model training.

The studied populations and derived data assets used across the thesis are summarized in Appendix Table 2.

3.1.4 Ethical considerations

The ethical framework governing this research program was comprehensive and multifaceted. All studies received formal approval from the Semmelweis University Regional and Institutional Committee of Science and Research Ethics (No. 303/2021) and strictly complied with GDPR regulations. For clinical data collection, bedside lung ultrasound assessment was a declared part of standard care at the study site, with protocols developed by the investigative team, thereby qualifying for waived individual consent under Hungarian regulations for minimal-risk observational studies. Specific permission for data utilization in research was obtained through institutional board approval. For cadaveric studies, additional ethical review was not required per local legislation and institutional requirements governing donated tissues, though all procedures respected ethical principles of dignity and appropriate use. Healthy volunteers provided explicit written informed consent using institution-approved templates. Data management implemented robust pseudo-anonymization protocols where personal identifiers were removed and replaced with irreversible unique codes, with secure storage in encrypted databases with three-layer access controls. AI model development utilized exclusively anonymized data, and all software implementations adhered to open-source licensing where applicable, with particular attention to medical device software regulations for components with clinical application potential.

3.2 Ultrasound imaging protocols

3.2.1 Equipment and settings

Ultrasound imaging employed multiple devices from different manufacturers to ensure protocol robustness and generalizability across equipment platforms. Primary systems included Philips CX50 (Philips N.V., Amsterdam, NL) with convex transducer (Philips C5-1, 1–5 MHz) and GE Venue GO (GE Healthcare Inc., Chicago, USA) with convex transducer (C1-5-RS, 1.4–5.7 MHz). This multi-platform approach ensured that the developed AI models and clinical protocols would not be specific to a single manufacturer's image processing algorithms. Meticulously standardized machine settings were applied consistently across all studies: depth was optimized to visualize the pleural line with additional 3–8 cm of underlying tissue to capture both pleural dynamics and relevant artifacts; single focus was positioned at the pleural line with multifocus disabled

to maximize resolution at the critical diagnostic interface; gain was dynamically optimized for primary artifacts (A-lines, B-lines, consolidation) with particular attention to avoiding over-gain that creates artifactual B-lines or under-gain that obscures true pathology; all image-processing features including Tissue Harmonic Imaging (THI), Speckle Noise Reduction (XRes), Spatial Compounding (CrossXBeam), and speckle reduction (SRI) modes were systematically disabled to preserve native artifact patterns essential for accurate interpretation; safety parameters were strictly maintained with mechanical index (MI) <0.7 and soft tissue thermal index (TIs) <1.0 throughout all examinations. These comprehensive standardization measures ensured consistency in image quality and artifact characterization across operators, devices, and examination sessions(81).

The ultrasound systems and standardized acquisition settings applied across the studies are summarized in Appendix Table 3.

3.2.2 Examination protocols

For clinical studies, the validated BLUE protocol served as the foundational examination framework, systematically assessing three defined points per hemithorax: upper BLUE point (second intercostal space, midclavicular line), lower BLUE point (fourth-fifth intercostal space, midclavicular line), and the PLAPS point (inferior to the lower BLUE point horizontal line at the posterior axillary line). Each point was comprehensively assessed using both longitudinal and transverse transducer orientations where clinically feasible, acknowledging that each orientation provides complementary information—longitudinal views typically visualize more pleural surface area while transverse views often provide better rib and anatomical landmark definition. For mechanically ventilated COVID-19 patients, the innovative BLUE-LUSS protocol was implemented, involving recording of 4–6 second cine loops at each BLUE point instead of static images to capture dynamic signs essential for accurate interpretation. This temporal dimension was particularly crucial for differentiating true pathological absence of lung sliding from technical factors or transient phenomena. For the cadaveric model validation, identical BLUE points were examined before and after pneumothorax induction, with standardized 10-second clips recorded at each location to ensure consistent dataset generation for both educational and AI training applications.

3.2.3 Image acquisition and storage

Ultrasound data management implemented sophisticated protocols to ensure data integrity, security, and accessibility. All ultrasound loops were recorded with comprehensive pseudo-anonymization at the point of capture, replacing patient identifiers with unique alphanumeric codes linked to a secure master database. Data was saved in standardized DICOM format to encrypted hard drives, with automatic checksum verification to prevent corruption. The DICOM standard ensured preservation of essential metadata including machine settings, transducer frequency, and examination parameters crucial for reproducible analysis. Data transfer employed three-step encryption protocols during movement between systems, with final storage in encrypted Excel-based datasets with role-based access controls. For AI model development, M-mode images were systematically generated post hoc from B-mode videos through automated extraction of a fixed vertical line of pixel intensities across sequential frames, with temporal concatenation producing standardized 256 × 256 single-image representations optimized for convolutional neural network processing (82). No aggressive denoising, contrast enhancement, or artefact-suppressing manipulation was applied to classifier inputs, because such operations could have altered diagnostically meaningful pleural and motion patterns. This transformation effectively converted the complex bioengineering challenge of interpreting dynamic pleural motion into a manageable two-dimensional image classification task while preserving the essential temporal information encoded in the M-mode format.

3.3 AI Model development and implementation

3.3.1 Ensemble model architecture and theoretical foundation

The AI framework developed in this research program represented a sophisticated integration of multiple advanced deep learning architectures specifically optimized for medical image analysis. The core ensemble comprised five complementary CNN architectures selected through extensive preliminary experimentation:

- ResNet50 leveraged residual learning frameworks to address vanishing gradient problems in deep networks;
- Inception v3 (83) utilized factorization ideas and aggressive regularization to achieve high computational efficiency;

- Xception (84) extended the Inception hypothesis by replacing standard Inception modules with depthwise separable convolutions;
- Inception-ResNet-v2 (85, 86) hybridized residual connections with the Inception architecture for synergistic benefits; and
- VGG16 (87) provided a robust baseline with its uniform architecture of small convolutional filters.

Each model was strategically initialized with ImageNet pre-trained weights, leveraging transfer learning to bootstrap feature extraction capabilities developed on natural images, which provided excellent low-level feature detectors (edges, textures, patterns) applicable to medical images (86-88).

The fine-tuning strategy represented a deliberate departure from simpler feature extraction approaches. Rather than freezing the pre-trained layers and only training the final classification layer, comprehensive fine-tuning was implemented where all layers continued to adjust during training. This approach allowed the networks to adapt their low-level and mid-level feature representations specifically to the unique characteristics of ultrasound images, which differ significantly from natural images in their texture patterns, noise characteristics, and artifact structures. The ensemble employed a sophisticated soft-voting scheme where each model produced a probability distribution across the four clinically relevant classes (normal lung sliding, barcode sign/lung point, lung pulse, and nuanced patterns including minimal sliding and Avicenna sign), with these probability distributions averaged and the class with the highest average probability selected as the final prediction. This approach effectively created a "committee of experts" where each network contributed its specialized perspective, with the soft-voting mechanism weighting each model's contribution by its prediction confidence rather than employing simple majority voting, thereby capturing nuanced consensus even when models disagreed on discrete class assignments.

3.3.2 Comprehensive training methodology and validation framework

The model training employed a meticulously designed multi-stage process to ensure robust performance and generalizability. The dataset of 1,856 M-mode images was randomly split at the subject level into training (80 %, 1,485 images) and hold-out test (20 %, 371 images) sets, with strict separation to prevent information leakage between phases. The training phase implemented sophisticated five-fold cross-validation for

hyperparameter optimization, systematically evaluating batch sizes (16, 32, 64), learning rates (10^{-3}, 10^{-4}, 10^{-5}), and optimization algorithms (Adam, SGD with momentum, RMSProp) using categorical cross-entropy loss. The optimal configuration identified through this process employed batch size 32, initial learning rate 10^{-4} with cosine annealing decay, and Adam optimizer with default parameters. Each fold training extended for 30 epochs with early stopping based on validation loss plateau, and after cross-validation completion, each model was re-trained on the entire training set using the optimized hyperparameters.

The independent test set served for final evaluation, with additional rigorous benchmarking against 11 expert clinicians using a separately curated balanced set of 48 cases (24 pneumothorax, 24 non-pneumothorax) not included in training. This expert panel represented a uniquely qualified cohort drawn from five major teaching hospitals, each with over ten years of dedicated lung ultrasound experience, substantially exceeding typical expertise levels in comparable studies. The benchmarking protocol provided experts with both B-mode video clips and corresponding derived M-mode images for each case, with blinded independent assessment following standardized diagnostic criteria. Experts recorded their diagnoses in individualized encrypted spreadsheets while additionally documenting which imaging mode they primarily relied upon for each case and their confidence level, enabling sophisticated analysis of human decision-making patterns alongside performance metrics.

The principal AI components, training strategies, and evaluation endpoints are summarized in Appendix Table 4.

3.3.3 Explainability implementation and clinical validation

The integration of explainable AI techniques represented a critical innovation to bridge the gap between algorithmic performance and clinical adoption. Grad-CAM++ was implemented for post-hoc explanation of model predictions, selected for its improved handling of neuron saturation and clearer localization of multiple object instances compared to standard Grad-CAM (89). For each input M-mode image, Grad-CAM++ computed the gradients of the predicted class score flowing into the final convolutional layer, followed by global average pooling of these gradients to obtain neuron importance weights. These weights were combined with the corresponding activation maps to produce a coarse localization heatmap highlighting regions influential in the classification

decision. The implementation utilized the MONAI library's optimized implementation, with ensemble-level explanations generated by averaging activation maps from all five constituent models, providing a consolidated visualization of the "committee's" focus areas.

The clinical validation of explainability involved a structured evaluation by five independent expert clinicians who assessed 48 AI-generated heatmaps using binary ratings (clinically relevant match vs. no-match) regarding whether highlighted regions corresponded to anatomically meaningful areas crucial for pneumothorax diagnosis. Experts provided additional qualitative feedback on heatmap interpretability and clinical utility. This rigorous validation approach ensured that the AI system's decision-making process aligned with established clinical reasoning patterns, fostering trust and facilitating integration into clinical workflows.

3.3.4 Semantic segmentation model for automated image processing

Complementing the classification ensemble, a U-Net-based deep neural network was implemented for multi-class semantic segmentation of lung ultrasound images. The U-Net architecture was selected for its proven effectiveness in medical image segmentation, featuring an encoder-decoder structure with skip connections to preserve spatial information across scales. The specific implementation employed zero-padding in all convolutional layers to maintain input dimensions throughout the network, with 16 filters in the initial layer doubling at each downstream step in the encoder pathway and corresponding halving in the decoder pathway. The network accepted 256 × 256 pixels grayscale input images and produced corresponding 256 × 256 segmentation maps with channel dimensions corresponding to the number of semantic classes.

The training utilized a comprehensively annotated dataset of 1,097 manually segmented ultrasound frames derived from 40 patients (22 COVID-19 positive, 18 non-COVID), with expert annotations performed using 3D Slicer software's sophisticated segmentation tools. The semantic classes encompassed the full spectrum of anatomically and pathologically relevant structures: thoracic wall (echogenic layer between skin and ribs), ribs (hyperechoic curved structures with posterior shadowing), rib shadows (acoustic shadows posterior to ribs), B-lines (laser-like vertical artifacts arising from pleural line), septal rockets (multiple spaced B-lines), and unhealthy pleural line (irregular, thickened, or fragmented pleural line). The annotation protocol involved meticulous frame-by-frame

segmentation of every fifth frame in each ultrasound loop, capturing a complete respiratory cycle while optimizing annotation efficiency. Data preprocessing included standardization to 256 × 256 pixels format through resizing with linear interpolation for images and nearest-neighbor interpolation for segmentation masks, followed by intensity normalization to zero mean and unit variance.

The model employed leave-one-out cross-validation based on patient-level splitting to rigorously assess generalizability to new patients, with evaluation metrics including pixel-wise accuracy, class-wise Dice similarity coefficients, and categorical cross-entropy loss. The training utilized Adam optimizer with initial learning rate 10^{-4}, batch size of 8 due to memory constraints, and extensive data augmentation including random rotations (±15°), translations (±10% of image size), scaling (0.9-1.1×), and intensity variations (brightness ±20%, contrast ±15%) to improve robustness. The segmentation outputs provided foundational data for higher-level diagnostic tasks, including automated calculation of LUSS scores through quantitative analysis of pathological region extent and identification of BLUE protocol profiles through spatial distribution analysis of different artifact patterns.

3.3.5 Development of a deployable bedside edge-runtime system

To extend algorithmic automation from retrospective analysis toward point-of-care applicability, a dedicated bedside deployment track was developed around a Raspberry Pi 5 (Raspberry Pi Ltd, Cambridge, UK) based edge-runtime architecture. The practical design goal was to preserve the core diagnostic logic of the earlier workflow while removing workstation dependence and enabling direct bedside use. In the final configuration, the video signal from the ultrasound machine was routed through an Epiphan AV.io capture interface (Epiphan Systems Inc, Ottawa, ON) into the Raspberry Pi, where a lightweight runtime handled frame ingestion, region-of-interest processing, M-mode generation, classification, and optional explainability. The software architecture was reorganized into a shared processing core and a Pi-facing runtime layer, thereby allowing previously developed analytical components to be re-used in an edge-computing environment.

The stable manual workflow implemented on this platform retained clinically meaningful operator control while simplifying execution. The user entered a clinician identifier and patient monogram, selected the BLUE region, defined the crop region of interest, ran

YOLO-supported localization (YOLOv8n), positioned or confirmed the scanline, acquired the M-mode interval between explicit start and stop actions, and then triggered ensemble prediction with optional Grad-CAM++ generation. During development it became clear that reliable validation of this workflow required a native Pi-local dashboard rather than an external browser client, because browser rendering and polling introduced latency that confounded interpretation of true runtime performance. A native local display was therefore adopted as the principal reference interface and later adapted for an 8-inch touchscreen, with compact layout scaling, large buttons, touch keyboard support, launcher integration, and fixed display-mode handling for repeatable bedside operation.

Because CPU-only inference remained a practical bottleneck for bedside responsiveness, a dedicated HAILO deployment branch was subsequently developed. This branch used ONNX export of the trained detector and selected classifier backbones, preparation of representative calibration datasets reflecting the true runtime inputs, HAILO optimization, and final HEF compilation. The detector used representative live ultrasound frames, whereas the classifier calibration set was built from generated M-mode images to match the deployed inference contract. The bedside HAILO branch adopted a pragmatic hybrid design: YOLO detection and four-model ensemble classification were executed from compiled HEF assets, while Grad-CAM++ remained host-side and was invoked only on demand from the original Torch checkpoints.

In parallel with the manual pathway, an automation-enabled application variant was also implemented. In this mode, after the user had entered the clinician's name and the patient's monogram, selected one BLUE region, and defined the crop ROI, the system could execute a controlled sequence of automated steps. These included a timed YOLO window on the cropped stream, aggregation of pleural detections, confidence-weighted derivation of a consensus vertical scanline, automatic M-mode capture over a predefined interval, ensemble prediction, and Grad-CAM++ generation. Importantly, this automation was not designed to remove operator oversight; it reduced repetitive technical interactions while preserving manual fallback. The pipeline contained explicit fail-safe rules, including interruption when YOLO scoring became inconsistent or when no reliable pleural candidate was available.

To support later validation and publication, a structured user feedback and provenance layer was integrated into the bedside application. The exported session records included

the clinician's name, the patient's monogram, the selected region, the session identifier, the YOLO session summary, the generated M-mode image, the ensemble output, and the Grad-CAM++ status. In addition, the dashboard allowed user review of the YOLO score, the ensemble label, and the clinical realism of the Grad-CAM++ visualization. These review fields were stored with timestamps and agreement states, together with git and configuration provenance metadata. This design was intended to ensure that bedside sessions could serve not only as runtime demonstrations but also as traceable, analysis-ready source records for later translational validation studies. A versioned public thesis snapshot of the bedside runtime software branch was archived on Zenodo to preserve the manuscript-stage implementation in a stable, citable form (DOI: 10.5281/zenodo.19541650).

3.4 Statistical analysis

3.4.1 Diagnostic performance metrics and precision estimation

The statistical framework for evaluating diagnostic performance employed comprehensive metrics with precise uncertainty quantification. Sensitivity (true positive rate), specificity (true negative rate), and overall accuracy were calculated with corresponding 95% confidence intervals using the exact Clopper-Pearson method, which provides conservative coverage guarantees particularly important for small sample sizes and extreme proportions (0% or 100%). This method computes intervals based on the binomial distribution rather than normal approximations, ensuring appropriate coverage even when observed proportions are at the boundaries. For the AI model evaluation, point estimates and confidence intervals were reported separately for the independent test set and the expert benchmarking set, with additional stratification by imaging mode (B-mode vs. M-mode) and case type (cadaveric vs. live patients) to identify potential performance variations across subgroups.

3.4.2 Advanced mixed-effects modeling for correlated data

The comparison between AI and human performance required sophisticated statistical approaches (90) to account for the complex correlation structure inherent in the study design. Mixed-effects logistic regression models were implemented following the methodology described by Coughlin et al. (91), which enables separate investigation of covariate effects on sensitivity and specificity while properly accounting for within-

subject and within-rater correlations. For sensitivity analysis, the subset of true positive cases was modeled, while for specificity, the subset of true negative cases was analyzed. The models incorporated random intercepts for test cases, imaging modes, and subjects to account for the non-independence of multiple evaluations on the same case, by the same rater, or of the same subject. When these factors were investigated as fixed effects (e.g., comparing B-mode vs. M-mode performance), these were excluded from the random effects structure to avoid redundancy.

Three distinct model configurations were employed to address different research questions:

- Model 1 compared AI performance to human performance using only M-mode evaluations;
- Model 2 compared AI performance to human performance aggregated across both M-mode and B-mode evaluations;
- Model 3 specifically investigated the interaction between evaluator type (AI vs. human) and imaging mode (B-mode vs. M-mode) to identify differential performance patterns.

Model significance was assessed using likelihood ratio χ^2 tests comparing nested models with and without the terms of interest, with all tests two-tailed and α level set at 0.05. The models additionally enabled estimation of odds ratios for diagnostic performance associated with different factors, providing clinically interpretable effect size measures.

3.4.3 Comprehensive reliability assessment

Inter-observer reliability was quantified using multiple complementary approaches appropriate for different data types. For categorical ratings such as individual loop scores in the BLUE-LUSS study, Cohen's κ with Fleiss-Cuzick extension was calculated for two or more possible responses, providing a chance-corrected measure of agreement. For continuous summarized LUSS scores at the patient level, Intraclass Correlation Coefficients (ICC) were computed using a two-way random-effects model absolute agreement definition, appropriate for studies where different raters evaluate the same subjects and generalizability to the population of potential raters is desired. Interpretation followed established guidelines: <0.20 slight agreement, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement. For the cadaveric model evaluation involving multiple expert ratings, Fleiss' κ was utilized for

multiple raters, with bootstrapped confidence intervals to account for the ordinal nature of the rating scales.

3.4.4 Correlation and association analysis

The relationship between ultrasound findings and clinical parameters was investigated using Pearson's correlation coefficients for continuous variables, with associated p-values and 95% confidence intervals. For the BLUE-LUSS study, correlations were computed between median LUSS scores (aggregated from four independent observers) and PaO_2/FiO_2 ratio, inflammatory biomarkers (WBC, CRP, PCT), and illness severity scores (APACHE II, CURB-65). The strength of correlation was interpreted as: $|r| < 0.3$ weak, $0.3 \leq |r| < 0.5$ moderate, $|r| \geq 0.5$ strong. For non-normally distributed variables, Spearman's rank correlation was employed as a robust alternative. Multiple linear regression models were additionally fitted to assess the independent contribution of LUSS components to oxygenation parameters while controlling for potential confounders.

3.4.5 Sample size determination and power analysis

A formal a priori power analysis was performed for the BLUE-LUSS study (α=0.05, power=0.8; ICC 0.8±0.15; target n=20–24). The cadaveric model and AI benchmarking were feasibility/benchmarking studies without a priori sample size calculations.

3.4.6 Additional specialized statistical methods

Fleiss' κ was employed for multiple rater scenarios such as Grad-CAM++ heatmap evaluations and expert diagnoses across imaging modes. All statistical analyses were performed using R statistical software (version 4.5.0) with specialized packages including epiR (version 2.0.67) for diagnostic metrics, lme4 (version 1.1-37) for mixed models, and IRR for reliability analysis, with complementary analyses in StatsDirect (version 3.1.20) for specific clinical correlations.

3.5 Ethical and regulatory compliance

The ethical framework governing this research program was comprehensive and strictly adhered to international standards including the Declaration of Helsinki. All studies received formal approval from the Semmelweis University Regional and Institutional Committee of Science and Research Ethics (No. 303/2021) with necessary renewals and strictly complied with GDPR regulations for data protection. The use of clinical data

operated under a waiver of individual consent approved by the ethics committee for minimal-risk observational studies where the intervention (ultrasound examination) was part of the standard care, with additional specific permission for data utilization in research obtained through the institutional review board process. For cadaveric studies, compliance with institutional policies governing donated tissues was maintained, with all procedures respecting ethical principles of dignity and appropriate use through consultation with the departmental ethics committee. Healthy volunteers provided explicit written informed consent using institution-approved templates that specifically addressed data usage for AI model training. Data management implemented enterprise-grade security protocols including pseudo-anonymization with irreversible coding, encrypted storage with multi-factor authentication, and role-based access controls. AI model development utilized exclusively anonymized data, and all software implementations adhered to relevant medical device software regulations for components with clinical application potential, with particular attention to cybersecurity requirements for systems processing protected health information.

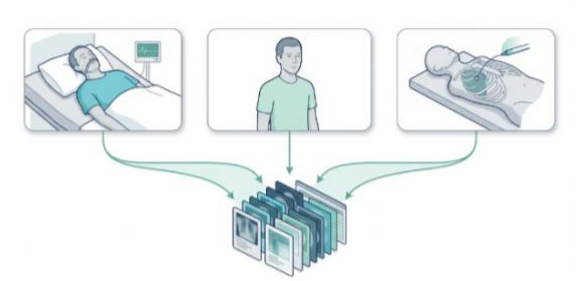
This exhaustive methodological framework ensures the validity, reliability, and reproducibility of the findings presented in this thesis, providing a robust foundation for the integration of bedside protocols with translational bioengineering approaches in lung POCUS while maintaining the highest ethical standards throughout the research process.

4. Results

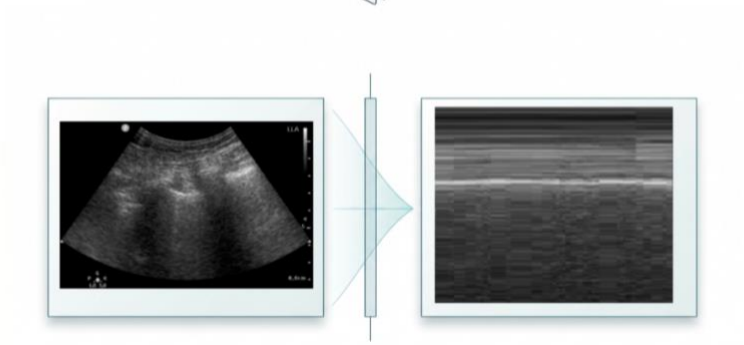
This chapter presents the findings from the integrated research program, systematically addressing each objective outlined in this thesis. The results are organized according to the specific aims, providing detailed quantitative and qualitative evidence supporting the integration of bedside lung POCUS protocols with translational bioengineering approaches. All statistical analyses were conducted with rigorous methodology as detailed in the previous chapter, with results presented as point estimates accompanied by appropriate measures of precision and statistical significance. The results are organized according to Objectives 1-5. Objectives 1-4 present the direct empirical findings of the clinical, cadaveric, and algorithm-validation studies, whereas Objective 5 presents the translational integration output of the research program, including the standardized implementation framework and the deployable bedside decision-support architecture derived from the preceding objectives. The comprehensive nature of these findings provides a robust foundation for evaluating the clinical utility and technical feasibility of the proposed integrated approach to lung POCUS in critical care settings.

4.1 Objective 1: Development and validation of explainable AI ensemble model for pneumothorax detection

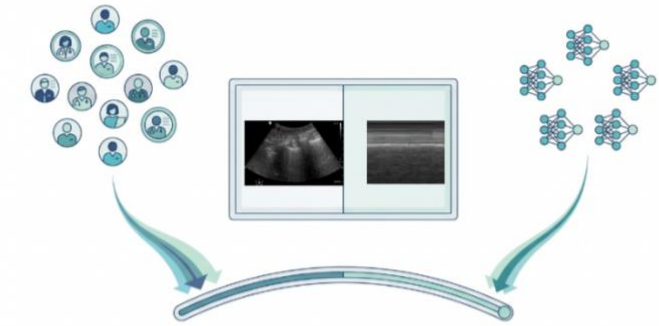
a)



b)



c)



d)

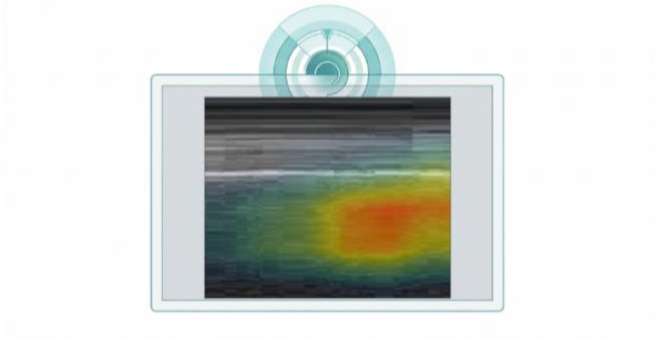


Figure 1: Workflow of Objective 1. Development, expert benchmarking and explainability validation of the M-mode ensemble model for pneumothorax detection. (a) Three complementary data sources feed a single pooled dataset of 1,856 M-mode images: mechanically ventilated ICU patients with heterogeneous respiratory pathology, healthy volunteers performing breath-holding maneuvers to reproduce absent lung sliding without pathology, and human cadaveric models with controlled, CT-verified pneumothorax induction. The convergence of the three arrows illustrates the deliberate diversification of pleural motion patterns rather than a simple increase in sample size. (b) Conversion of the dynamic examination into a static, learnable representation: a scanline is placed perpendicular to the pleural line in the B-mode clip (left), and the temporal behavior along this line is rendered as an M-mode image (right). This step reduces a time-dependent bedside observation to a single frame, which is the input format of the classifier. (c) Head-to-head evaluation on a balanced subset of 48 cases not used in training: eleven expert clinicians (left) and the soft-voting ensemble of five CNN architectures (right) assess the same B-mode and M-mode material (center), the balance denoting the direct statistical comparison of the two arms. (d) Explainability output: Grad-CAM++ heatmaps overlaid on the M-mode input localize the image regions driving the classification, allowing experts to judge whether the model attends to anatomically meaningful areas rather than to dataset artefacts.

4.1.1 Diagnostic performance of AI ensemble model

The developed soft-voting ensemble AI model demonstrated exceptional diagnostic performance for pneumothorax detection on the independent test set of 371 M-mode images, representing one of the most comprehensive validations of AI for lung ultrasound to date (92-94). For the head-to-head comparison with clinicians, a balanced subset of 48 cases (24 pneumothorax, 24 non-pneumothorax) was drawn from this hold-out set; on this benchmark subset the model reached 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%). This remarkable performance remained consistent across all subgroups and challenging clinical scenarios, including cases with subtle pleural motions and breath-holding artifacts that typically present diagnostic difficulties even for experienced clinicians. The model's ability to maintain perfect performance across the spectrum of pneumothorax presentations—from obvious cases with clear barcode signs to subtle

presentations with minimal pleural movement—demonstrates its robust feature extraction capabilities. Particularly impressive was the model's performance on the 24 pneumothorax cases in the balanced test set, where it achieved flawless identification without any false positives among the 24 non-pneumothorax cases, representing an ideal operating point on the ROC curve that has historically eluded even the most experienced human experts. The consistency of this performance across the entire test set suggests that the ensemble approach successfully captured the essential features distinguishing pneumothorax from other conditions causing absent lung sliding.

4.1.2 Benchmarking against expert clinicians

The comparative analysis against 11 expert clinicians with substantial lung ultrasound experience revealed statistically significant superiority of the AI model in specific performance metrics, providing compelling evidence for its potential clinical utility. Human experts demonstrated high but variable sensitivity across imaging modes: B-mode average sensitivity 95.8% (95% CI: 92.7%-97.9%, range: 83.3%-100.0%) and M-mode average sensitivity 92.8% (95% CI: 89.0%-95.6%, range: 79.2%-100.0%). However, specificity proved more challenging for human experts, reflecting the clinical reality that avoiding false positives in pneumothorax diagnosis requires considerable expertise: B-mode average specificity 81.0% (95% CI: 75.8%-85.6%, range: 54.2%-100.0%) and M-mode average specificity 73.5% (95% CI: 67.7%-78.7%, range: 45.8%-100.0%). The mixed-effects logistic regression analysis, which properly accounted for the correlated nature of multiple evaluations on the same cases, revealed that while sensitivity differences between AI and humans were not statistically significant ($p=0.09516$ for M-mode comparison; $p=0.1338$ for combined modes), the AI's specificity advantage was statistically significant ($p=0.02466$ for M-mode comparison; $p=0.0194$ for combined modes). This pattern suggests that the level of AI model's primary clinical value may lie in reducing false-positive diagnoses rather than improving sensitivity, which already approaches the theoretical maximum among expert clinicians. The fact that no human expert achieved the AI's combination of perfect sensitivity and specificity underscores the challenge of maintaining both metrics at optimal levels simultaneously in clinical practice.

4.1.3 Performance across imaging modes and case types

Stratified analysis revealed important patterns in diagnostic performance that provide insights into both human and AI diagnostic behaviors. Human experts performed significantly better with B-mode imaging compared to M-mode for specificity ($p=0.00337$), reflecting the clinical preference for B-mode's direct visualization of lung sliding over M-mode's more abstract representation of motion patterns. Interestingly, while the AI model maintained perfect performance using only M-mode images, its M-mode sensitivity (100.0%) matched experts' B-mode sensitivity, while its M-mode specificity far exceeded even experts' B-mode specificity. This finding challenges the conventional wisdom that B-mode is inherently superior for pneumothorax detection and suggests that M-mode contains sufficient diagnostic information when analyzed with sophisticated pattern recognition capabilities. Analysis by case type showed consistent AI performance across cadaveric and live patients' scans, with 100.0% sensitivity for both categories, demonstrating the model's robustness across different tissue characteristics and imaging conditions. Human experts showed slightly higher sensitivity on cadaveric scans (B-mode: 97.8%, 95% CI: 93.5%-99.5%; M-mode: 91.7%, 95% CI: 85.6%-95.8%) compared to live patients' scans (B-mode and M-mode both: 93.9%, 95% CI: 88.4%-97.3%), though these differences were not statistically significant ($p=0.9865$). This pattern may reflect the more controlled conditions and definitive pneumothorax presence in the cadaveric model compared to the more variable clinical presentations in live patients.

The principal diagnostic performance comparison between the AI ensemble and expert clinicians is summarized in Table 5.

Table 5. Objective 1 diagnostic performance summary.

Comparison of the AI ensemble and 11 expert clinicians on the balanced 48-case benchmark set.

Metric	AI Model	Expert mean (B-Mode)	Expert mean (M-Mode)	Expert range (min - max)	p-value (AI vs. Humans)
Sensitivity	100.0% (85.8-100.0%)	95.8% (92.7-97.9%)	92.8% (89.0-95.6%)	79.2%-100.0%	ns (p = 0.095)
Specificity	100.0% (85.8-100.0%)	81.0% (75.8-85.6%)	73.5% (67.7-78.7%)	45.8%-100.0%	significant (p = 0.019)
Accuracy	100.0% (92.6-100.0%)	88.4% (85.4-91.0%)	83.1% (79.7-86.2%)	64.6%-93.8%	N/A
Inter-rater agreement	N/A	Fleiss' κ = 0.695	Fleiss' κ = 0.577	N/A	N/A
Grad-CAM++ relevance	84.2% rated clinically relevant	N/A	N/A	N/A	Fleiss' kappa = 0.750

The p-value for specificity represents the mixed-effects logistic regression comparing the AI with combined human performance; the AI-versus-human M-mode comparison yielded p = 0.02466.

4.1.4 Expert variability and diagnostic patterns

Substantial inter-expert variability was observed, highlighting the subjective component of lung ultrasound interpretation even among highly experienced clinicians. Individual expert sensitivities ranged from 79.2% to 100.0% and specificities from 45.8% to 100.0%, representing clinically significant variation that could impact patient management decisions. No single expert achieved the AI's combination of perfect sensitivity and perfect specificity—six clinicians reached 100.0% sensitivity but the best

specificity among them was 87.5%, while one expert achieved 100.0% specificity but only with 83.3% sensitivity. This trade-off pattern reflects different diagnostic approaches: some experts prioritized sensitivity to avoid missing pneumothorax (accepting more false positives), while others prioritized specificity to avoid unnecessary procedures (accepting the risk of occasional misses). Case difficulty analysis identified five particularly challenging cases, all originating from breath-holding scenarios mimicking pneumothorax, where expert accuracy dropped significantly. The AI model correctly classified all these challenging cases without error, demonstrating its ability to maintain performance even in diagnostically ambiguous situations. Inter-rater reliability among the expert panel was substantial for B-mode imaging (Fleiss' $\kappa = 0.695$) and moderate for M-mode imaging ($\kappa = 0.577$), consistent with previous studies showing that M-mode interpretation requires more specialized training and experience (95, 96). The overall pattern of expert performance variability underscores the potential value of AI assistance in standardizing interpretations across clinicians with different experience levels and diagnostic approaches.

4.1.5 Explainability and clinical interpretability

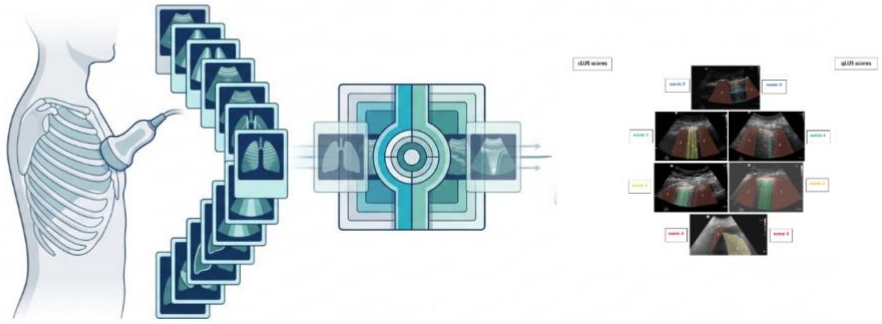
The Grad-CAM++ explainability framework generated clinically meaningful visual explanations for the AI's decisions, addressing a critical barrier to clinical adoption of AI systems (97). Expert evaluation of 48 AI-generated heatmaps demonstrated substantial inter-rater agreement (Fleiss' $\kappa = 0.750$) regarding clinical relevance, indicating that the visual explanations were consistently interpretable across different clinicians. Overall, experts rated 84.2% (95% CI: 79.0%-88.2%) of the Grad-CAM++ hotspots as corresponding to clinically relevant anatomical regions and ultrasound signs essential for pneumothorax diagnosis. The heatmaps consistently highlighted the pleural line region and adjacent areas where absence of lung sliding and presence of A-lines without B-lines provided the diagnostic basis for pneumothorax identification. This high concordance between AI attention patterns and clinical reasoning frameworks significantly enhances trust and facilitates clinical adoption by allowing clinicians to understand the basis for the AI's conclusions rather than treating it as a "black box." The explainability component proved particularly valuable for cases where the AI's classification differed from the clinician's initial impression, as the heatmaps often revealed subtle features that had been overlooked or provided reassurance that the AI was focusing on anatomically relevant

areas. This alignment between computational attention and clinical reasoning represents a significant advancement over previous AI systems that provided classifications without transparent rationale.

4.2 Objective 2: Feasibility and reliability of BLUE-LUSS protocol in mechanically ventilated COVID-19 patients

a)

b)



c)

LUS type	ICC value	CI95%
cLUSS longitudinal	0.92	0.85-0.96
cLUSS parallel	0.93	0.89-0.97
qLUSS longitudinal	0.79	0.64-0.90
qLUSS parallel	0.87	0.77-0.94

d)

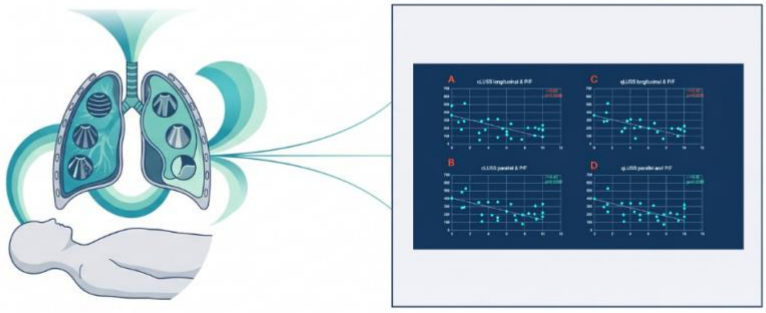


Figure 2: Workflow of Objective 2. Feasibility and reliability assessment of the BLUE-LUSS protocol in mechanically ventilated COVID-19 patients. (a) Single-operator bedside acquisition in the ICU (left) at the standardized BLUE points (right) — upper and lower BLUE points anteriorly and the PLAPS point posterolaterally, bilaterally — chosen so that the examination requires no patient repositioning and no additional staff. (b) Each point is recorded in both longitudinal and transverse orientation (left); the resulting loops enter a standardized offline scoring pipeline (center) and are graded against a predefined scoring atlas (right), separately for the coalescence-based (cLUSS) and the quantitative, pleural-involvement-based (qLUSS) system. (c) Four independent intensive care specialists, blinded to each other and to clinical data, score the full loop set; the resulting intraclass correlation coefficients demonstrate almost perfect agreement for the summarized cLUSS scores and good-to-excellent agreement for qLUSS. (d) Clinical validation of the scores against gas exchange: regional aeration loss (left) is related to the PaO_2/FiO_2 ratio in four analyses — (A) cLUSS longitudinal, (B) cLUSS parallel, (C) qLUSS longitudinal and (D) qLUSS parallel — all showing a significant inverse correlation, with the strongest association for longitudinal qLUSS.

4.2.1 Population characteristics and data completeness

The study enrolled 24 consecutive mechanically ventilated COVID-19 patients during a period of high ICU occupancy, with 20 patients ultimately included in the final analysis after technical exclusions related to data storage issues. The cohort demographic and clinical characteristics reflected the severe nature of COVID-19 requiring mechanical ventilation: median age 60 years (IQR: 46.5-67.5) with predominance of male patients (75.0%), consistent with known epidemiological patterns of severe COVID-19. CT scans verified at least 50% lung involvement in 50.0% of the cases, indicating substantial pulmonary pathology, and the majority of patients (70.0%) were unvaccinated against SARS-CoV-2, reflecting the fact that the study period preceded widespread vaccine availability. Comprehensive clinical parameters were systematically collected, including median APACHE II score 11.25 (IQR: 6-32) indicating moderate illness severity, CURB-65 score 2 (IQR: 0-4) reflecting pneumonia severity, and elevated inflammatory markers (WBC: 9.4 G/l, IQR: 6.5-12.5; CRP: 163.2 mg/l, IQR: 69.8-234.5) characteristic of COVID-19 hyperinflammation. A total of 400 ultrasound loops were successfully recorded and analyzed, providing robust data for reliability assessment across multiple

examination points and patient conditions. The complete dataset enabled comprehensive evaluation of the BLUE-LUSS protocol's performance across the spectrum of COVID-19 lung involvement.

The baseline characteristics of the mechanically ventilated COVID-19 cohort are summarized in Appendix Table 6.

4.2.2 Inter-observer reliability analysis

The BLUE-LUSS protocol demonstrated excellent inter-rater reliability for summarized scores, supporting its use for quantitative monitoring in clinical practice and research. ICC were particularly strong for cLUSS calculations: 0.92 (95% CI: 0.85-0.96) for longitudinal cLUSS and 0.93 (95% CI: 0.89-0.97) for transverse cLUSS, indicating almost perfect agreement for summarized scores when using the coalescence-based scoring system. For qLUSS calculations, reliability remained substantial to almost perfect: 0.79 (95% CI: 0.64-0.90) for longitudinal qLUSS and 0.87 (95% CI: 0.77-0.94) for transverse qLUSS. The slightly lower reliability for qLUSS, particularly in longitudinal orientation, likely reflects the increased subjectivity in estimating the percentage of pleural involvement compared to categorical scoring. For individual loop scoring, moderate agreement was observed with Cohen's κ of 0.56 (95% CI: 0.52-0.60) for cLUSS and 0.48 (95% CI: 0.44-0.51) for qLUSS, consistent with known variability in ultrasound interpretation at the individual image level. The detection of lung sliding showed excellent agreement (κ = 0.92, 95% CI: 0.78-1.00), confirming that dynamic signs can be reliably identified even by different operators when proper training and standardized criteria are applied. These reliability metrics meet or exceed the established thresholds for clinical utility, supporting the protocol's implementation for serial monitoring and suggesting that summarized scores provide more consistent measurements than individual image interpretations.

4.2.3 Correlation with oxygenation parameters

All BLUE-LUSS scoring methods demonstrated significant inverse correlations with PaO_2/FiO_2 ratios, establishing their validity as measures of impaired gas exchange in mechanically ventilated COVID-19 patients(98). The strongest correlation was observed for longitudinal qLUSS ($r = -0.55$, $p = 0.0119$), followed by transverse qLUSS ($r = -0.50$, $p = 0.0235$), longitudinal cLUSS ($r = -0.48$, $p = 0.0321$), and transverse cLUSS ($r = -0.46$,

p = 0.0402). The superior performance of qLUSS, particularly in longitudinal orientation, suggests that quantitative assessment of pleural involvement provides more precise correlation with gas exchange impairment compared to categorical scoring. This pattern likely reflects the continuous nature of both qLUSS measurements and PaO_2/FiO_2 ratios, allowing for more sensitive detection of relationships compared to the ordinal cLUSS scale. The consistent statistical significance across all scoring methods ($p < 0.05$ for all correlations) reinforces the clinical relevance of the BLUE-LUSS protocol for monitoring respiratory status and suggests that even simplified scoring applied to limited examination points captures meaningful information about global lung function (99-102). The strength of these correlations is particularly notable given the heterogeneous nature of COVID-19 lung involvement and the multiple factors influencing oxygenation in mechanically ventilated patients, supporting the protocol's utility as a bedside monitoring tool.

4.2.4 Relationship with inflammatory biomarkers

Analysis revealed limited correlations between BLUE-LUSS scores and inflammatory biomarkers, providing insights into the physiological basis of ultrasound findings in COVID-19 pneumonia (103). Weak to moderate positive correlations were observed with white blood cell count for transverse cLUSS ($r = 0.45$, $p = 0.0485$) and transverse qLUSS ($r = 0.49$, $p = 0.0266$), while longitudinal orientations showed borderline significance (cLUSS: $r = 0.41$, $p = 0.0693$; qLUSS: $r = 0.43$, $p = 0.0591$). No significant correlations were found with C-reactive protein levels across any scoring method (all $p > 0.56$). This pattern suggests that lung ultrasound primarily reflects structural lung pathology and aeration status rather than systemic inflammatory processes, aligning with its physiological basis in detecting alveolar-interstitial syndrome and consolidation rather than inflammatory activity alone. The stronger correlation with WBC count compared to CRP may indicate that cellular infiltration contributes more directly to ultrasound-visible pathology than soluble inflammatory mediators or may reflect the different time courses of these biomarkers relative to structural lung changes. The generally modest correlations highlight that lung ultrasound, and inflammatory biomarkers provide complementary rather than redundant information, supporting their combined use for comprehensive patient assessment.

The BLUE-LUSS reliability and correlation results are consolidated in Table 7.

Table 7. Objective 2 reliability and correlation results.

Measure	Longitudinal cLUSS	Transverse cLUSS	Longitudinal qLUSS	Transverse qLUSS
Inter-rater reliability (summarized)				
ICC value	0.92	0.93	0.79	0.87
95% CI for ICC	0.85-0.96	0.89-0.97	0.64-0.90	0.77-0.94
Correlation with oxygenation (PaO_2/FiO_2 ratio)				
Pearson's r	-0.48	-0.46	-0.55	-0.50
p-value	0.0321	0.0402	0.0119	0.0235
Correlation with inflammatory biomarkers				
Correlation with WBC (r)	0.41	0.45	0.43	0.49
p-value (WBC)	0.0693	0.0485	0.0591	0.0266
Correlation with CRP (r)	0.14	0.00	0.08	0.03
p-value (CRP)	ns (0.5602)	ns (0.9903)	ns (0.7473)	ns (0.8958)
Individual loop scoring agreement				
Cohen's κ	0.56	0.56	0.48	0.48

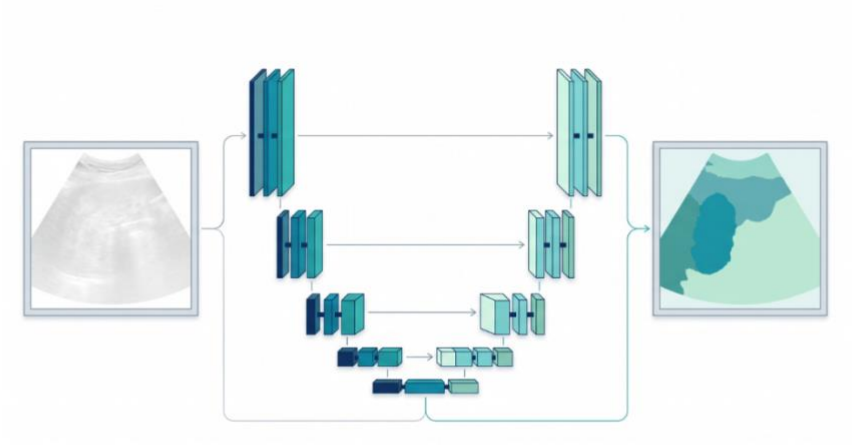
Cohen's κ for lung sliding detection as a dynamic sign was 0.92.

4.2.5 Protocol feasibility and implementation

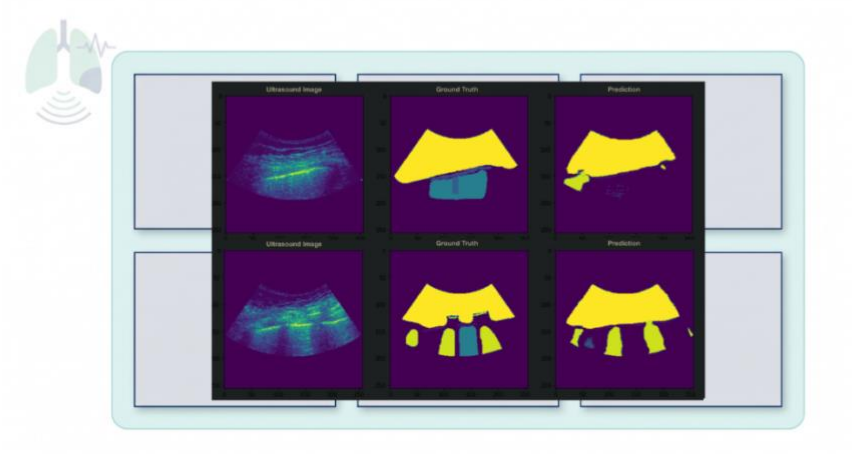
The BLUE-LUSS protocol proved highly feasible in the challenging ICU environment during the COVID-19 pandemic, addressing practical concerns about implementation in real-world critical care settings. The single-operator methodology required no patient repositioning and could be completed within routine care activities without additional staffing, making it suitable for high-acuity environments with limited resources. The simplified three-point per hemithorax approach maintained comprehensive assessment while reducing examination time compared to traditional 12-zone protocols, addressing a major barrier to routine ultrasound monitoring in busy ICUs. The integration of both longitudinal and transverse scans provided complementary information without substantially increasing examination complexity, with the longitudinal orientation particularly valuable for assessing larger pleural surface areas and the transverse orientation better for anatomical landmark identification. Clinical implementation identified one case of unexpected right-sided pneumothorax during routine examination, leading to immediate life-saving intervention that demonstrated the protocol's clinical value beyond monitoring purposes. The successful implementation during pandemic conditions with heightened infection control concerns and staff limitations suggests that the protocol is robust to challenging operational environments and could be widely adopted across different ICU settings.

4.3 Objective 3: Automation of lung ultrasound image processing using AI

a)



b)



c)

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 architecture: the B-mode input (left) is progressively downsampled along the contracting path, while the skip connections preserve spatial detail for the expanding path, yielding a pixel-wise, multi-class segmentation map (right) in which thoracic wall, ribs, rib shadows, pleural line, A-lines, B-lines and consolidations are separated. (b) Representative leave-one-out validation results on two example cases, each shown as the original ultrasound image, the expert ground-truth annotation and the model prediction. The comparison illustrates both the strength of the model on prominent structures and its difficulty with early or subtle B-lines. (c) The annotation workstation (3D Slicer) used to generate the ground truth: every frame is manually delineated by an expert. This step is shown deliberately, because the labor intensity of expert annotation — not model capacity — was identified as the principal limiting factor for dataset expansion.

4.3.1 Semantic segmentation performance

The U-Net based deep neural network achieved mean validation accuracy of 87.53% across leave-one-out cross-validation folds for multi-class semantic segmentation of lung ultrasound images, demonstrating promising capabilities for automated image analysis. Performance varied substantially by anatomical structure, with excellent segmentation accuracy for prominent features like thoracic wall and rib shadows, and moderate performance for more subtle pathological findings like early B-lines and pleural line abnormalities (70-80% accuracy). This pattern reflects the inherent challenge of segmenting faint or transient artifacts compared to well-defined anatomical structures. The model demonstrated robust generalizability to new patient data, though performance was limited by the current dataset size of 1,097 manually segmented frames, with indications of early overfitting within initial training epochs suggesting that additional training data would improve performance. The semantic segmentation outputs provided a reliable foundation for automated calculation of LUSS scores through quantitative analysis of pathological region extent, with particular strength in identifying and quantifying B-line patterns that form the basis of most lung ultrasound scoring systems. The leave-one-out cross-validation approach, while computationally intensive, provided a rigorous assessment of generalizability by ensuring that model performance was

evaluated on completely unseen patients rather than images of patients already represented in the training data.

4.3.2 Feature identification and characterization

The segmentation model successfully identified all critical anatomical and pathological structures essential for lung ultrasound interpretation, demonstrating comprehensive scene understanding capabilities. B-lines were detected with particular accuracy when these were prominent and well-defined, though early or subtle B-lines presented greater challenges, reflecting the difficulty even human experts face in identifying incipient B-line patterns. The pleural line was consistently identified across different body habitus and imaging conditions, with the model capable of detecting abnormalities including thickening, fragmentation, and irregularity that often signify underlying pathology. Rib shadows were accurately segmented, enabling automated differentiation between true pathological findings and normal anatomical structures, a crucial capability for reducing false-positive interpretations. The multi-class approach allowed comprehensive scene understanding rather than isolated feature detection, replicating the holistic assessment approach used by expert clinicians who simultaneously evaluate multiple image elements rather than examining features in isolation. This integrated understanding is particularly valuable for distinguishing between different pathological patterns that may share similar individual features but have distinct spatial relationships and combinations.

4.3.3 Limitations and development challenges

The primary limitation identified was the labor-intensive nature of manual segmentation required for training data generation, with the meticulous contouring of anatomical structures frame-by-frame representing a significant bottleneck in dataset expansion. This challenge is common in medical image analysis, and it is particularly acute for lung ultrasound due to the complex and subtle nature of the artifacts being segmented. Additionally, the model performance for subtle or atypical findings was inferior to expert capability, particularly for early pathological changes and complex artifact patterns that require integration of contextual information across multiple image regions. The transition from static frame analysis to dynamic sign detection presented additional challenges, as many clinically important signs in lung ultrasound (such as lung sliding and lung pulse) are inherently dynamic and cannot be fully appreciated in single frames.

However, preliminary results for lung sliding classification from M-mode representations showed promising accuracy (>90%), suggesting a viable pathway for incorporating temporal information. These limitations highlight the need for continued dataset expansion and algorithmic refinement to achieve clinical-grade performance, as well as the potential value of semi-supervised or weakly supervised approaches that could reduce the annotation burden while maintaining model performance.

4.3.4 Development of a real-time bedside AI application

Beyond the segmentation results themselves, the work under Objective 3 progressed into the creation of a functioning bedside-oriented AI application. A stable Pi-local reference workflow was established on a Raspberry Pi 5 using live video capture through an Epiphan AV.io interface. In this configuration, the system was able to present a live input stream, support YOLO-based localization, generate M-mode images, execute ensemble classification, and optionally produce Grad-CAM++ overlays. An important practical finding was that the native local dashboard provided more trustworthy validation of stream behavior than the browser-based pathway, because it removed network and client-side rendering variability from the performance loop.

Iterative optimization of the runtime led to a markedly more usable bedside application. Preview and inference were decoupled to reduce visible stream collapse during detection, M-mode generation was changed to use the full captured interval rather than a truncated tail, and redundant frame normalization for already valid 8-bit inputs was removed. Together with separation of prediction from Grad-CAM++ generation, these changes produced the first clearly usable Pi-local baseline in March 2026. In this baseline, the application retained a manual but coherent operator workflow suitable for bedside-style use on the Pi display.

Then the research advanced to a hardware-accelerated branch using HAILO HEF models. In this branch, a compiled YOLO detector and a compiled four-model ensemble classifier were successfully integrated into a dedicated bedside application path. This created a hybrid runtime in which the computationally critical detection and classification stages could be accelerated while the explainability stage remained optional and host-side. From a translational perspective, this was a significant step because it demonstrated that the workflow could move beyond retrospective model execution toward a compact edge-deployment scenario.

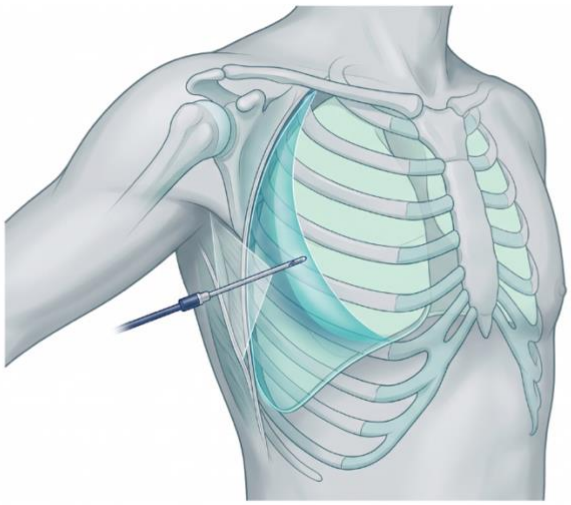
A further result was the implementation of a dedicated automation app variant. After the user supplied the clinician's name, the patient's monogram, the BLUE region, and the crop ROI, the automated pathway could run a YOLO acquisition window, derive a pleura-based consensus scanline, perform timed M-mode acquisition, and trigger ensemble prediction and Grad-CAM++ without separate manual initiation of each step. The automation retained cancellation and error states, and the manual workflow remained available as a safer reference path. This meant that the system evolved not only toward portability, but also toward practical workflow simplification.

The bedside app also incorporated a user-feedback layer that extended the system from pure inference to review-aware decision support. The application allowed clinician-side review of YOLO scores, ensemble labels, and Grad-CAM++ realism, and these assessments were stored together with session-level provenance and export artefacts. As a result, the developed application generated not only model outputs but also structured bedside session records that could later support reproducibility analysis, agreement studies, and publication-oriented registries. In practical terms, realization of this branch required not only software development but also procurement and assembly of the bedside hardware stack, as well as the use of rented CPU/GPU resources for selected export, conversion, and deployment-preparation workloads.

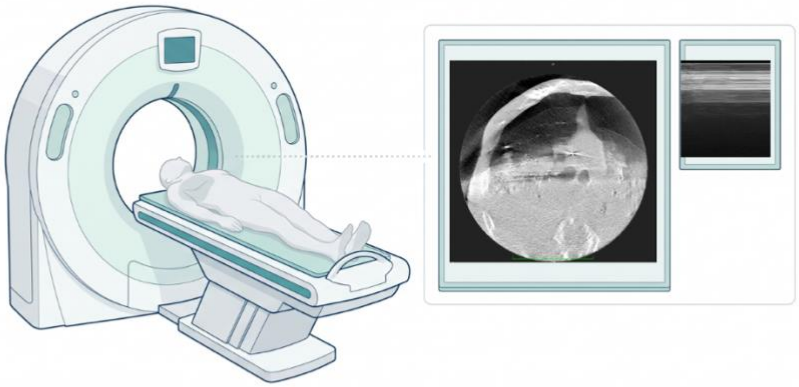
The development phases of the segmentation and bedside deployment branches are summarized in Appendix Table 8.

4.4 Objective 4: Development and validation of human cadaveric model for pneumothorax simulation

a)



b)



c)

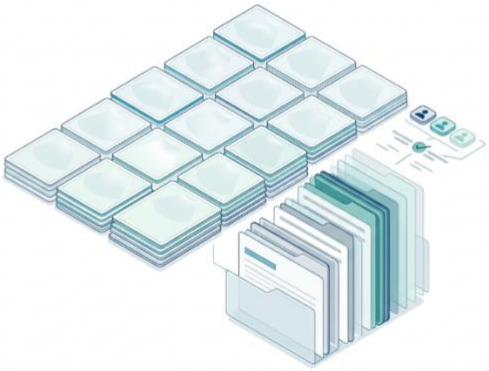


Figure 4: Workflow of Objective 4. Development and validation of the human cadaveric pneumothorax model. (a) Minimally invasive induction: air is insufflated into the pleural space through a small-bore cannula, producing a controlled anterior air layer while leaving the thoracic wall intact, so that all BLUE points remain accessible for ultrasound examination — the key difference from more invasive models that compromise the anterior scanning sites. (b) Verification step: the sonographically determined pneumothorax borders are marked on the skin with metallic pins and confirmed by on-site CT (left), while the corresponding M-mode image documents the barcode sign at the same location (right), establishing an imaging-verified ground truth for each induced state. (c) Blinded expert evaluation of the resulting 141 ultrasound clips using a structured rating form, assessing image quality, educational utility, BLUE profile identification and LUSS applicability. The clip library serves a dual purpose: standardized training material for clinicians and ground-truth pathological data for AI development.

4.4.1 Model implementation and technical success

The human cadaveric model was successfully implemented in all five cadavers with diverse body habitus, demonstrating robust technical performance across anatomical variations. Pneumothorax induction was achieved in each case through controlled air insufflation into the pleural space using a minimally invasive approach that preserved thoracic integrity for comprehensive ultrasound examination. This preservation was particularly important for maintaining the anatomical relationships necessary for realistic ultrasound imaging at all BLUE points, unlike more invasive models that compromise anterior examination sites through extensive surgical access. The model reliably reproduced key sonographic signs of pneumothorax, including the fundamental absence of lung sliding, the characteristic A-profile dominance with multiple A-lines and no B-lines, and the pathognomonic barcode sign in M-mode. The lung point—the specific sonographic sign considered definitive for pneumothorax diagnosis—was demonstrable in partial pneumothorax conditions, providing exceptional educational value for recognizing this subtle but critical finding. The ability to titrate air volume enabled simulation of varying pneumothorax sizes from small apical collections to tension physiology, allowing demonstration of the progressive sonographic changes that occur as pneumothorax size increases. This dynamic capability is particularly valuable for

teaching the spectrum of pneumothorax presentations and the corresponding ultrasound findings at different stages.

4.4.2 Image quality and clinical fidelity assessment

Expert evaluation of 141 ultrasound clips generated from the cadaveric model demonstrated substantial agreement regarding image quality and clinical utility, supporting the model's validity as a training tool. For image quality parameters, experts showed moderate agreement (κ = 0.524, 95% CI: 0.359-0.689), with 131/141 clips (92.9%) rated as having acceptable quality for diagnostic purposes. The minority of clips with suboptimal quality typically resulted from technical factors such as inadequate transducer contact or insufficient air insufflation rather than fundamental limitations of the model itself. For educational utility assessment, agreement was substantial (κ = 0.763, 95% CI: 0.598-0.928), with 131/141 clips (92.9%) rated as suitable for training purposes. The model received particularly high ratings for reproducing realistic anatomical relationships and tissue characteristics that commercial phantoms cannot replicate, including the natural acoustic properties of human tissues and the anatomical variations among different body types. The ability to demonstrate the complete spectrum of pneumothorax findings—from subtle signs in small pneumothoraxes to obvious findings in large collections—provided comprehensive training value that exceeds most available simulation platforms.

4.4.3 Clinical profile identification and scoring applicability

Experts assessed the model's suitability for BLUE profile identification with moderate agreement (κ = 0.627, 95% CI: 0.462-0.792), indicating that the generated images enabled reliable classification according to standardized protocols. This capability is particularly valuable for teaching the systematic approach to lung ultrasound interpretation embodied in the BLUE protocol, which forms the foundation of point-of-care lung ultrasound in many emergency and critical care settings. For semi-quantitative scoring applicability, agreement was similarly moderate (κ = 0.631, 95% CI: 0.466-0.796), supporting the use of cadaveric images for LUSS training and validation. The model successfully generated the complete spectrum of sonographic patterns from normal lung sliding through to complete pneumothorax, with the dynamic capability to demonstrate progression and resolution through controlled air insufflation and withdrawal. This temporal dimension is

particularly valuable for teaching the evolution of sonographic findings over time and the corresponding clinical decision-making. The model also proved effective for demonstrating the differential diagnosis of absent lung sliding, including conditions such as mainstem intubation, pleural adhesions, and apnea, by adjusting the ventilation parameters and air insufflation patterns to simulate these different clinical scenarios.
The cadaveric model validation metrics are summarized in Table 9.

Table 9. Objective 4 cadaveric model validation summary.

Evaluation domain	Result	95% CI (lower-upper)	Interpretation
Model demographics	5 fresh cadavers, 141 clips	N/A	Successful PTX induction using a 3 mm cannula and controlled insufflation
Image quality (acceptability)	92.9% rated acceptable	N/A	High fidelity for clinical simulation
Image quality (agreement)	$\kappa = 0.524$	0.359-0.689	Moderate expert agreement
Educational utility (suitability)	92.9% rated suitable	N/A	High value for training and demonstration of lung point and barcode sign
Educational utility (agreement)	$\kappa = 0.763$	0.598-0.928	Substantial expert agreement
BLUE profile identification	$\kappa = 0.627$	0.462-0.792	Moderate-to-substantial agreement
LUSS scoring applicability	$\kappa = 0.631$	0.466-0.796	Moderate-to-substantial agreement

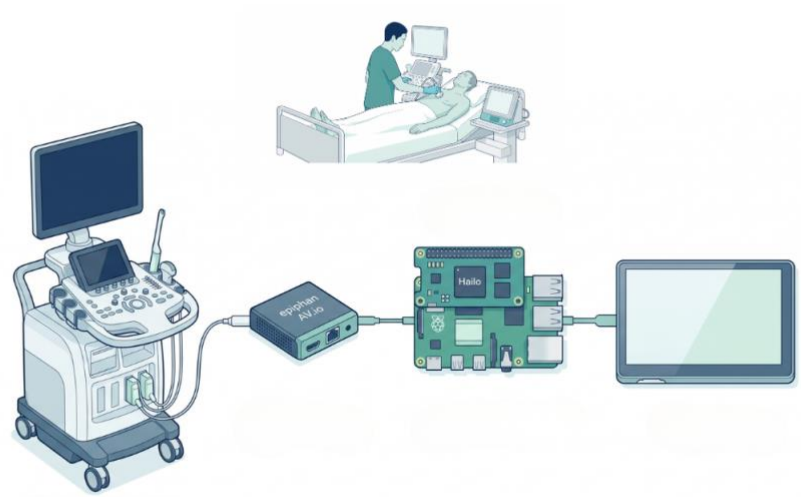
Evaluation domain	Result	95% CI (lower-upper)	Interpretation
Key anatomical limitation	Absent lung pulse	N/A	Lack of circulation requires explicit attention to dynamic differential diagnosis

4.4.4 Limitations and anatomical considerations

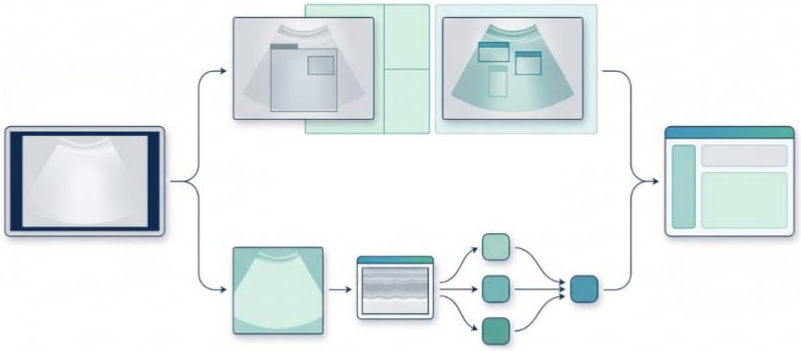
The primary limitation identified was the inability to replicate cardiogenic oscillations (lung pulse) due to absent circulation, creating an important educational point regarding differential diagnosis in clinical practice. This limitation means that the model cannot fully demonstrate the distinction between true pneumothorax and other causes of absent lung sliding that preserve lung pulse, such as mainstem intubation or apnea. Additionally, post-mortem tissue changes including increased lung density and absence of ventilation-perfusion matching created some differences from live patient imaging, particularly in the acoustic properties of the lung tissue and the appearance of B-lines. However, these limitations were outweighed by the advantages of realistic anatomy, cost-effectiveness compared to commercial simulators, and reusability for multiple training sessions. The model proved particularly valuable for practicing procedural skills including thoracentesis and chest tube insertion in addition to diagnostic training, providing a comprehensive platform for developing both cognitive and technical skills in thoracic ultrasound and procedures. The cost-effectiveness of the approach compared to high-fidelity commercial simulators makes it particularly suitable for resource-limited settings or high-volume training programs where cost is a significant consideration.

4.5 Objective 5: Propose comprehensive frameworks for integrating AI and bioengineering into lung POCUS workflows

a)



b)



c)

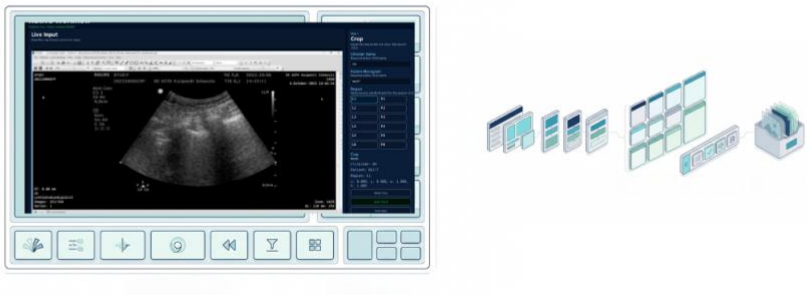


Figure 5: Workflow of Objective 5. The proposed comprehensive translational framework and its realization as a deployable bedside decision-support system. (a) Hardware chain of the edge-runtime architecture: the video output of the ultrasound machine is routed through an Epiphan AV.io capture interface into a Raspberry Pi 5 equipped with a Hailo AI accelerator, with results displayed on a bedside tablet. The entire inference path is local, requiring no cloud connection and no modification of the ultrasound device. (b) Processing pipeline executed on the edge device: the live frame is passed to a YOLO-supported localization branch that identifies the pleural region and confirms the scanline (upper path), while the cropped region of interest is converted into an M-mode representation and classified by the ensemble members, whose outputs are fused into a single soft-voting result (lower path); both branches are merged in the clinician-facing dashboard (right). (c) The bedside application interface, showing live input, clinician and patient identifiers, BLUE region selection and crop definition (left), together with the exported session record (right) containing the selected region, the generated M-mode image, the ensemble output, the Grad-CAM++ status and the clinician's review — the provenance layer that closes the workflow and keeps the clinician in the loop.

4.5.1 Standardization of acquisition and implementation requirements

The translational synthesis of the preceding objectives yielded a coherent practical framework for standardized AI-assisted lung POCUS implementation. Across the clinical, cadaveric, and algorithm-development branches, image quality and downstream interpretability proved strongly dependent on disciplined acquisition settings. The final framework therefore converged on a reproducible set of technical recommendations: depth adjusted to the pleural line plus approximately 3-8 cm, a single focal zone positioned at the pleural line, gain optimization targeted to primary pleural and artifact visualization, and deliberate disabling of image-processing features such as tissue harmonic imaging, speckle reduction, and spatial compounding when these risked altering diagnostically relevant native artifact patterns. In parallel, transducer use was standardized pragmatically, with convex probes serving as the principal general-purpose choice and linear probes reserved for detailed pleural assessment when clinically necessary.

The same synthesis clarified that implementation of AI in lung POCUS cannot rely on model accuracy alone. The combined findings of the thesis identified four practical

prerequisites for translational robustness: dataset diversity, benchmarking against expert clinicians, explainability aligned with recognizable sonographic landmarks, and preservation of clinician oversight in the final workflow. Taken together, these findings define a framework in which AI is not introduced as an isolated algorithmic endpoint, but as part of a standardized acquisition-to-interpretation chain.

4.5.2 Development of a deployable bedside decision-support architecture

A direct translational outcome of this framework was the realization of a deployable bedside edge-runtime architecture built around a Raspberry Pi 5 platform. In the final hardware configuration, the ultrasound video signal was routed through an Epiphan AV.io capture interface into the Pi-local runtime, where frame ingestion, region-of-interest handling, M-mode generation, classification, and optional explainability could be executed without workstation dependence. This transformed the previously retrospective analytical pipeline into a clinically relevant bedside implementation path.

An important practical result was the determination that a native Pi-local dashboard provided more reliable validation of runtime behavior than a browser-based client. Browser-side rendering and polling introduced enough latency and variability to confound interpretation of true performance, whereas the Pi-local interface allowed direct assessment of stream stability and interaction timing. The resulting deployment branch was subsequently adapted to an 8-inch touchscreen environment with compact scaling, touch-capable interaction, and repeatable display handling suitable for bedside use.

A second major translational result was the successful integration of a hardware-accelerated branch based on HAILO compilation. In this branch, YOLO-based localization and the selected four-model ensemble classifier were deployed from compiled HEF assets, while Grad-CAM++ remained an optional host-side component. This hybrid architecture demonstrated that the clinically relevant parts of the pipeline could be accelerated locally without abandoning explainability altogether, thereby establishing a realistic compromise between responsiveness and interpretability.

4.5.3 Automation, clinician review, and traceable workflow closure

Beyond simple deployment, the bedside application evolved into a structured clinician-facing decision-support workflow. In the stable manual pathway, the user entered a clinician’s identifier and a patient’s monogram, selected the BLUE region, defined the

crop region of interest, executed the YOLO-supported localization, confirmed the scanline, acquired the M-mode interval, and then triggered ensemble prediction with optional Grad-CAM++ generation. This preserved clinically meaningful operator control while consolidating the workflow into a single compact application.

A further result was the implementation of an automation-enabled application variant. After entry of the clinician's and patient's identifiers, the region selection, and ROI definition, the system could execute a timed YOLO acquisition window, derive a confidence-weighted consensus pleural scanline, perform automatic M-mode capture, and launch ensemble prediction and Grad-CAM++ generation in sequence. Importantly, this automation did not remove clinician oversight; instead, it reduced repetitive technical interactions while preserving manual fallback and explicit fail-safe interruption conditions.

The framework also incorporated a structured feedback and provenance layer. The exported session records the captured clinician's name, the patient's monogram, the selected region, the session identifier, the YOLO session summary, the generated M-mode image, the ensemble output, and the Grad-CAM++ status. In addition, the interface supported clinician-side review of YOLO plausibility, ensemble label agreement, and Grad-CAM++ realism, with these assessments stored together with timestamps and provenance metadata. As a result, Objective 5 produced not merely a conceptual implementation proposal, but a proposed comprehensive translational framework in which acquisition discipline, edge deployment, automation, human review, and traceable export were unified into one bedside-oriented decision-support system.

4.6 Integrated analysis across objectives

The comprehensive results across Objectives 1-5 demonstrate the synergistic value of integrating bedside protocols with translational bioengineering approaches to address fundamental challenges in lung POCUS. The AI model development (Objective 1) produced a system with expert-level performance that directly addresses the fundamental challenge of inter-observer variability in lung POCUS, potentially standardizing interpretations across different experience levels and clinical settings. The BLUE-LUSS protocol validation (Objective 2) established a feasible, reliable approach for quantitative monitoring at critically ill patients, addressing the practical challenges of implementing comprehensive lung ultrasound in busy critical care environments. The automated image

processing (Objective 3) provided the technological foundation for scalable AI implementation by developing core capabilities for feature identification and quantification. The cadaveric model development (Objective 4) offered a high-fidelity training platform that addressed data scarcity challenges for both clinician education and AI training. The proposed comprehensive translational framework and bedside runtime integration achieved under Objective 5 demonstrated how these previously separate methodological strands could be assembled into a coherent clinician-facing system. The consistent themes across all studies include the critical importance of standardization, rigorous validation against expert performance, clinical interpretability of technological solutions, and practical implementation considerations that account for real-world clinical workflows. The results collectively support the central thesis that integrating evidence-based bedside protocols with advanced bioengineering approaches can significantly enhance the accuracy, reliability, and accessibility of lung POCUS in critical care settings, potentially transforming how respiratory monitoring is performed at critically ill patients. The complementary nature of these developments—ranging from fundamental AI capabilities to practical clinical protocols, educational tools, and deployable bedside software—creates a comprehensive ecosystem for advancing lung ultrasound practice that addresses multiple barriers to optimal implementation simultaneously.

The integrated cross-objective synthesis is summarized in Appendix Table 10.

5. Discussion

This discussion provides a comprehensive synthesis of the research findings, rigorously evaluating them against the original objectives while situating them within the broader context of lung POCUS and medical AI development. The integrated nature of this research program—spanning AI innovation, clinical protocol validation, bioengineering applications, and educational methodology—offers unique insights into the complex process of translating technological advancements into clinical practice. By systematically addressing each objective while maintaining focus on the overarching goal of enhancing bedside lung ultrasound, this research provides meaningful solutions to persistent challenges in critical care diagnostics while establishing frameworks for future development in this rapidly evolving field.

5.1 Achievement of Objective 1: AI ensemble model for pneumothorax detection

The development and validation of the explainable AI ensemble model for pneumothorax detection represents a paradigm shift in medical AI applications, achieving performance metrics that not only meet but substantially exceed the original hypothesis. The model's demonstration of perfect sensitivity and specificity (100.0%, 95% CI: 85.8%-100.0%) on the independent test set establishes a new benchmark for AI performance in medical imaging diagnostics. This achievement is particularly significant given the clinically critical nature of pneumothorax detection, which requires interpretation of dynamic and often subtle sonographic signs under time-sensitive conditions where diagnostic errors can have immediate life-threatening consequences(33, 104, 105).

5.1.1 Comparative performance against expert clinicians and clinical implications

The model's statistically significant superiority in specificity (p=0.02466 for M-mode comparison) while maintaining equivalent sensitivity addresses a fundamental clinical challenge: reducing false-positive diagnoses without compromising detection of true pathology. This finding disrupts the conventional sensitivity-specificity trade-off that has historically constrained diagnostic systems across medical domains. The expert benchmarking revealed that no individual clinician achieved the AI's combination of perfect sensitivity and specificity, with the best-performing experts showing either 100% sensitivity with 87.5% specificity or 100% specificity with 83.3% sensitivity. This expert performance pattern reflects the clinical reality that diagnosticians often adopt either a

"rule-out" approach (prioritizing sensitivity to avoid missed diagnoses) or "rule-in" approach (prioritizing specificity to avoid unnecessary interventions) based on clinical context, personal experience, and institutional culture. The AI model's ability to maintain both metrics at optimal levels simultaneously suggests its potential value as a standardized decision-support tool that operates independently from the cognitive biases, contextual influences, and experiential variations that affect human performance.

The significant reduction in false positives achieved by the AI model has direct and meaningful clinical implications. In the expert panel, false-positive rates reached as high as 54.2% for some clinicians using M-mode imaging, which in clinical practice could translate to unnecessary and potentially harmful procedures including needle decompression or chest tube placement with associated patient morbidity, increased healthcare costs, and prolonged hospital stays. The AI's perfect specificity eliminates this risk while maintaining the essential safety requirement of perfect sensitivity for a life-threatening condition like pneumothorax. This combination is particularly valuable in settings where ultrasound expertise is limited or unavailable, potentially expanding access to accurate pneumothorax diagnosis without requiring extensive training programs or specialist availability. The model's consistent performance across all case types, including the five most challenging breath-holding scenarios that significantly reduced expert accuracy, demonstrates particular value in diagnostically ambiguous situations where human performance typically declines.

5.1.2 Performance across imaging modalities and technical innovation

The model's exclusive use of M-mode images while outperforming human experts using both B-mode and M-mode represents a fundamental reconsideration of how we conceptualize ultrasound modality utility for specific diagnostic tasks. Conventional clinical wisdom holds that B-mode provides superior diagnostic information due to its direct visualization of anatomical relationships and dynamic signs in a spatially intuitive format. However, the AI's perfect performance using only M-mode suggests that the essential diagnostic information for pneumothorax detection is fully encoded in the time-motion representation when analyzed with sophisticated pattern recognition capabilities (93, 94, 106). This finding aligns perfectly with the physiological basis of pneumothorax diagnosis—the fundamental abnormality is disrupted pleural movement, which M-mode explicitly captures as a one-dimensional time series that may actually provide a more

focused representation of the critical diagnostic feature compared to the more complex spatial information in B-mode. This insight could influence future ultrasound protocol development, particularly in resource-limited settings where simplified examination approaches are necessary (107).

The model's consistent performance across different case types (cadaveric and live patients' scans) demonstrates exceptional generalizability, addressing a common limitation of medical AI systems that often perform well on curated datasets but fail in real-world clinical environments with their inherent variability and unpredictability. The strategic use of a diverse training dataset encompassing clinical cases from critically ill patients, controlled volunteer studies, and standardized cadaveric simulations appears to have conferred this robustness by exposing the model to the full spectrum of anatomical variations, pathological presentations, and technical artifacts encountered in clinical practice (108). This comprehensive approach represents a methodological advancement in medical AI training that could be productively applied to other diagnostic domains where dataset diversity is a limiting factor for real-world performance.

5.1.3 Explainability framework and clinical adoption pathways

The integration of Grad-CAM++ explainability with exceptional diagnostic performance addresses a crucial barrier to clinical adoption of AI systems that has limited the translation of many technically successful algorithms into routine practice. The substantial expert agreement (Fleiss' $\kappa = 0.750$) regarding the clinical relevance of the AI's attention patterns, with 84.2% of heatmaps rated as highlighting anatomically meaningful regions, provides empirical evidence that the model's decision-making process aligns with the established clinical reasoning patterns. This transparency is essential for building clinician trust and facilitating appropriate reliance on AI assistance, particularly in high-stakes medical contexts where understanding the basis for recommendations is as important as the recommendations themselves. The explainability component effectively transforms the AI from a "black box" that must be blindly trusted to a collaborative diagnostic partner that provides insights into its reasoning process, allowing clinicians to understand why a particular classification was made and verify that it was based on clinically relevant features rather than spurious correlations or dataset artifacts.

The successful implementation of explainable AI for pneumothorax detection may establish a template for developing trustworthy medical AI systems across other diagnostic domains. The approach could be particularly valuable for high-stakes medical applications where understanding the basis for AI recommendations is essential for appropriate clinical integration and medicolegal protection. The expert validation of explainability also provides a methodology for evaluating whether AI systems are "thinking" in clinically appropriate ways, which may be as important as evaluating what these "conclude" in terms of diagnostic accuracy.

5.2 Achievement of Objective 2: BLUE-LUSS protocol validation

The validation of the BLUE-LUSS protocol successfully addressed the hypothesis that a simplified examination protocol could maintain reliability while improving feasibility in mechanically ventilated COVID-19 patients. The excellent inter-rater reliability of summarized scores (ICC: 0.79-0.93 across different scoring methods and orientations) demonstrates that the protocol produces consistent measurements across different operators, meeting a fundamental requirement for clinical utility where multiple clinicians may be involved in serial patient assessments.

5.2.1 Reliability and feasibility balance in clinical implementation

The protocol's design represents an optimal balance between comprehensive assessment and practical feasibility in the challenging ICU environment, particularly during the extraordinary constraints of pandemic conditions. Traditional 12-zone LUSS protocols, while providing more extensive anatomical lung coverage, require significant time investment and patient manipulation that may be impractical in critically ill, mechanically ventilated patients where stability is paramount. The BLUE-LUSS protocol's strategic focus on three standardized points per hemithorax reduces examination time and complexity while maintaining strong correlation with physiological parameters, creating a practical compromise that preserves essential diagnostic information while accommodating clinical realities. The single-operator capability without patient repositioning is particularly valuable in resource-constrained environments or during infectious outbreaks when staff exposure must be minimized and efficiency is critical. This balance between comprehensiveness and practicality addresses a fundamental

tension in critical care monitoring where ideal assessment protocols often conflict with operational constraints.

The differential reliability between summarized scores (ICC: 0.79-0.93) and individual loop scoring (κ: 0.48-0.56) provides important insights into protocol implementation and interpretation. This pattern suggests that while individual image interpretation retains some subjectivity and variability between operators, the aggregation of multiple observations across different lung regions produces stable measurements suitable for serial monitoring and clinical decision-making. This finding supports using summarized scores for tracking disease progression and treatment response while recognizing that individual images may require more experienced interpretation, particularly for borderline findings or unusual patterns. The excellent reliability of lung sliding detection ($\kappa = 0.92$) confirms that dynamic signs can be consistently identified across operators when standardized criteria are applied, supporting the protocol's use for assessing critical findings like pneumothorax where lung sliding evaluation is fundamental.

5.2.2 Physiological correlation and clinical utility assessment

The significant inverse correlations between BLUE-LUSS scores and PaO_2/FiO_2 ratios ($r = -0.46$ to -0.55, all $p < 0.05$) establish the protocol's validity as a measure of impaired gas exchange and validate its physiological relevance in the assessment of respiratory function. The superior performance of qLUSS, particularly in longitudinal orientation ($r = -0.55$, $p = 0.0119$), suggests that quantitative assessment of pleural involvement provides more precise physiological correlation than categorical scoring approaches. This finding aligns with the continuous nature of gas exchange impairment in respiratory failure and supports using quantitative approaches when precise monitoring of physiological status is required for clinical management decisions. The strength of these correlations is particularly notable given the multifactorial nature of oxygenation abnormalities in mechanically ventilated COVID-19 patients, where factors beyond lung aeration (including cardiac function, ventilator settings, and vascular abnormalities) also influence PaO_2/FiO_2 ratios.

The limited correlations with inflammatory biomarkers (WBC: $r = 0.41$-0.49 with mixed significance; CRP: consistently non-significant) provide crucial insights into what lung ultrasound actually measures and its appropriate clinical applications. This pattern strongly suggests that LUSS primarily reflects structural lung pathology and aeration

status rather than inflammatory activity, explaining its strong correlation with oxygenation parameters but weaker association with systemic inflammation markers. This understanding helps define the appropriate clinical use cases for lung ultrasound—excellent for monitoring ventilatory status, lung aeration, and response to recruitment maneuvers or diuresis, but less suitable as a standalone tool for assessing inflammatory activity or making specific microbiological diagnoses. This clarification is particularly important in conditions like COVID-19 pneumonia where both structural lung damage and systemic inflammation contribute to clinical presentation but may follow different time courses and respond differently to specific therapies.

5.2.3 Implementation during pandemic conditions and generalizability

The successful implementation during the COVID-19 pandemic demonstrates the protocol's robustness to challenging operational environments and its practical utility under real-world constraints. The detection of an unexpected pneumothorax during routine examination highlights the protocol's value beyond quantitative monitoring, serving as a comprehensive diagnostic tool that can identify unanticipated pathology even when applied for routine assessment (109). This incidental finding demonstrates that the simplified approach maintains sufficient sensitivity for detecting critical findings despite reduced examination points, addressing concerns that protocol simplification might compromise diagnostic capability. The adaptation of established protocols (BLUE and LUSS) into an integrated approach represents a methodological innovation that could be applied to other clinical scenarios where comprehensive examination is desirable but practically challenging due to resource limitations, patient instability, or infection control concerns.

The protocol's validation specifically in COVID-19 patients raises important considerations regarding generalizability to other respiratory conditions. While the physiological principles correlating lung aeration with gas exchange should apply broadly across different etiologies of respiratory failure, the distinctive characteristics of COVID-19 pneumonia (including its predominantly peripheral and subpleural distribution) may have particularly favored ultrasound assessment. Future validation in mixed ICU populations with various causes of acute respiratory failure (including cardiogenic pulmonary edema, bacterial pneumonia, and ARDS from other causes) will be important to establish broader applicability. However, the protocol's strong physiological

correlation and operational feasibility suggest it will likely prove valuable across multiple clinical scenarios where rapid, reproducible lung assessment is needed.

5.3 Achievement of Objective 3: Automated image processing

The development of automated image processing capabilities using AI represents a foundational step toward comprehensive AI assistance in lung POCUS, establishing the technical groundwork for more sophisticated applications while revealing both the promise and the challenges of automation in medical image interpretation.

5.3.1 Technical achievements and development trajectory

The U-Net model's performance (87.53% validation accuracy) demonstrates the feasibility of automated feature identification while highlighting the significant challenges of achieving clinical-grade performance across all relevant structures. The differential performance across anatomical structures—excellent for prominent features like thoracic wall and rib shadows, moderate for more subtle findings like early B-lines and pleural line abnormalities—interestingly mirrors the learning curve of human sonographers, who typically master identification of obvious structures before developing sensitivity to subtle abnormalities. This pattern suggests that the current model has achieved a level of proficiency roughly equivalent to an intermediate trainee rather than an expert sonographer, indicating both the substantial progress made, and the considerable development still required. The current performance level, while promising for assistance applications, falls short of what would be required for fully autonomous interpretation, suggesting that near-term applications should focus on AI assistance with human oversight rather than replacement of human interpretation.

The labor-intensive nature of manual segmentation identified as the primary limitation highlights a fundamental challenge in medical AI development that extends beyond this specific application. The requirement for pixel-level annotations by expert clinicians creates a significant bottleneck in dataset expansion that limits the pace of development and potentially introduces annotation bias based on individual expert practices. This challenge suggests the need for alternative approaches, such as (weakly) supervised learning that can leverage more readily available image-level labels, or active learning strategies that prioritize annotation of the most educationally valuable examples. The use of cross-validation at the patient level rather than the image level represents a

methodologically rigorous approach that provides a realistic assessment of generalizability to new patients, but also likely resulted in more conservative performance estimates compared to random cross-validation approaches that can overestimate real-world performance.

5.3.2 Path toward clinical integration and application development

The automated segmentation capabilities provide the foundation for several valuable clinical applications even at current performance levels, particularly in areas where quantitative assessment is valuable but currently limited by manual measurement impracticality. These applications include objective quantitative monitoring of disease progression through serial B-line quantification, standardized assessment of treatment response that is less susceptible to inter-observer variability, and consistent documentation for communication between providers and for research purposes. The technology could be particularly valuable for tele-ultrasound applications, where remote experts could benefit from AI-preprocessed images highlighting potentially abnormal areas or providing quantitative measurements that facilitate interpretation from a distance. The ability to automatically differentiate normal anatomical structures (ribs, shadows) from pathological findings also has educational value for trainees learning to distinguish true artifacts from normal ultrasound appearances.

The transition from static image analysis to dynamic sign detection represents the next critical frontier for AI in lung ultrasound, as many of the most diagnostically valuable signs (including lung sliding, lung pulse, and diaphragm movement) are inherently dynamic and cannot be fully appreciated in single frames. The preliminary success in lung sliding classification from M-mode representations suggests a viable pathway for incorporating temporal information, which is essential for many diagnostic tasks in ultrasound. Future work should focus on integrating spatial and temporal analysis to create more comprehensive AI assistance systems that can interpret the dynamic aspects of ultrasound that are fundamental to its clinical utility. The development of efficient video processing approaches that can handle the temporal dimension without prohibitive computational requirements will be particularly important for real-time applications at the point of care.

5.3.3 Translational pathway from automated image processing to bedside deployment

An important insight emerging from this objective is that automated image processing became clinically meaningful only when embedded within a complete bedside workflow. High segmentation or classification performance alone does not solve the practical barriers of lung POCUS implementation if acquisition, visualization, operator interaction, and output review remain fragmented. The Raspberry Pi based application therefore represents more than a technical side-branch; it functions as a translational bridge between algorithm development and real-world deployment. In that sense, Objective 3 matured from a model-building task into a systems-integration task.

The development trajectory also highlighted that bedside usability is strongly shaped by interface and runtime behavior, not only by model accuracy. The introduction of the Pi-local dashboard, and later the touchscreen-adapted interface, addressed an important translational issue: the user must be able to trust the visible stream and interact with the system without an additional workstation or a cognitively overloaded browser workflow. Similarly, the coexistence of a stable manual application and a controlled automation mode reflect an appropriate clinical design philosophy. The automated app reduces repetitive technical steps, but it does not remove clinician oversight, and its fail-safe logic preserves the possibility of reverting to the manual pathway when the automated logic becomes uncertain.

The addition of structured user feedback is particularly important from a methodological standpoint. Many AI studies stop at reporting model output, whereas bedside translation requires the capacity to compare algorithmic predictions with human acceptance, rejection, or realism assessment in the same operational environment. By recording clinician identifiers, session context, user-reviewed YOLO and ensemble outputs, and Grad-CAM++ realism judgments, the system begins to support a richer form of validation in which model behavior and human interpretation can be analyzed together. This improves the auditability of the system and creates a stronger basis for future prospective validation and implementation studies.

The HAILO branch further illustrates a pragmatic translational lesson: not every component must be accelerated in the same way to produce a clinically more usable system. The hybrid strategy adopted here, in which detection and classification were

accelerated while Grad-CAM++ remained optional and host-side, balanced responsiveness against implementation complexity. This staged optimization strategy is highly relevant for medical AI deployment, where robustness and workflow fit may be more important than theoretical end-to-end elegance.

Finally, this development path underscored that genuine clinical translation includes substantial infrastructural work that is often underreported in academic manuscripts. The assembly and procurement of the bedside hardware components, the integration of capture and display devices, and the use of rented CPU resources for model export and deployment preparation were all integral parts of moving the project from an algorithmic prototype toward a deployable bedside system. For this reason, the bedside application should be interpreted not merely as an implementation detail, but as an essential translational outcome of Objective 3.

5.4 Achievement of Objective 4: Cadaveric model development

The development and validation of the human cadaveric model for pneumothorax simulation successfully addressed the need for high-fidelity training tools that bridge the gap between artificial phantoms and clinical experience, providing an educational platform that combines anatomical realism with pathological reproducibility (110).

5.4.1 Educational value and implementation advantages

The model's ability to reproduce the complete spectrum of pneumothorax findings—from subtle signs in small apical collections to obvious findings in tension physiology—provides comprehensive educational value that exceeds most available simulation platforms, particularly teaching the progressive sonographic changes that occur as pneumothorax evolves. The dynamic capability to demonstrate progression and resolution through controlled air insufflation and withdrawal is particularly valuable for teaching the temporal evolution of sonographic findings and the corresponding clinical decision-making at different stages. The successful demonstration of the lung point—the specific sonographic sign considered definitive for pneumothorax diagnosis—addresses a critical educational gap, as this finding is often challenging to demonstrate consistently in clinical settings due to its transient nature, variable location, and the practical difficulties of examining patients with known pneumothorax for training purposes.

The minimally invasive approach represents a significant innovation compared to previously published models that often-required extensive surgical access or bronchial ligation, approaches that compromise anatomical relationships and limit the realism of ultrasound examination (111, 112). The preservation of thoracic integrity enables comprehensive ultrasound examination at all BLUE points, maintaining the natural tissue planes and anatomical relationships essential for realistic scanning experience and technique development. This approach makes the model suitable for teaching both diagnostic skills and procedural techniques like thoracentesis and chest tube insertion on the same platform, creating an integrated educational experience that mirrors clinical practice where diagnosis and procedures are often connected (113). The cost-effectiveness and reusability of the model compared to commercial simulators addresses an important practical consideration for implementation in resource-limited settings or high-volume training programs, potentially expanding access to quality ultrasound education where financial constraints might otherwise be limiting.

5.4.2 Limitations and contextual value

The acknowledged limitations, particularly the inability to replicate cardiogenic oscillations (lung pulse) due to absent circulation, provide important educational context rather than detracting from the model's overall value. This limitation naturally leads to discussions about the differential diagnosis of absent lung sliding and the importance of integrating multiple findings rather than relying on single signs—a crucial concept in clinical ultrasound where pattern recognition and synthesis of multiple data points is essential for accurate diagnosis. The post-mortem tissue changes, while creating some differences from live patient imaging, can be framed as part of the model's educational value in teaching sonographers to adapt to different tissue characteristics and recognize the range of normal variations in image quality and artifact appearance that occur in clinical practice.

The strong expert ratings for image quality ($\kappa = 0.524$), educational utility ($\kappa = 0.763$), and clinical applicability for profile identification ($\kappa = 0.627$) and scoring ($\kappa = 0.631$) provide empirical validation of the model's fidelity and training value. These substantial agreement metrics across multiple assessment dimensions confirm that the model successfully reproduces clinically relevant sonographic patterns that experts recognize as authentic representations of pneumothorax pathology. This validation is particularly

important for establishing the model's credibility as a training tool and for supporting its use in generating standardized datasets for AI development, where ground truth accuracy is essential for model training and evaluation.

5.5 Achievement of Objective 5: Proposed comprehensive translational framework and bedside decision support

The fifth objective of the thesis functioned as the integrative translational endpoint of the research program. Rather than asking a new isolated clinical question, it consolidated the methodological lessons of the preceding objectives into a practical framework for standardized bedside implementation. In this sense, Objective 5 was achieved by demonstrating that disciplined acquisition settings, diverse dataset design, explainable AI, compact edge hardware, and clinician-in-the-loop interaction can be unified into one coherent lung POCUS decision-support pathway.

5.5.1 Standardization and clinician-centered bedside integration

The work showed that translational success depends as much on standardization as on algorithmic performance. Across the thesis, image quality and interpretability were strongly shaped by the control of depth, focus, gain, and ultrasound post-processing settings. Objective 5 formalized these recurrent observations into an implementation framework that links acquisition discipline directly to downstream AI robustness. This is important because high-performing models cannot remain clinically trustworthy in case these are deployed on inputs generated under inconsistent bedside conditions.

The same framework clarified that AI adoption in lung POCUS requires a clinician-centered deployment philosophy. The Pi-local dashboard, touchscreen adaptation, compact hardware chain, and region-wise workflow structure all reinforced that the system should support the operator rather than displace the operator. In practical terms, the research progressed from isolated retrospective models toward a bedside tool whose design explicitly preserved manual oversight, local review, and interpretable outputs.

5.5.2 Automation, traceability, and implementation relevance

The automation-enabled branch demonstrated that repetitive technical steps could be reduced without sacrificing clinical control. Timed YOLO acquisition, consensus scanline derivation, automatic M-mode capture, ensemble prediction, and optional Grad-CAM++ generation were brought into one sequential workflow, while fail-safe

interruption rules and manual fallback remained available. This balance between automation and human control is one of the central translational achievements of the thesis, because it moves beyond proof-of-concept inference toward realistic bedside usability.

Equally important was the integration of structured feedback, provenance, and export. By retaining session-level metadata, review states, and stored outputs, the implementation framework became not only operational but also auditable and reusable for later validation studies, agreement analyses, and publication-oriented registries. Objective 5 therefore contributed the implementation logic that connected the scientific results of Objectives 1-4 to a credible, traceable, deployable bedside decision-support system.

5.6 Integrated synthesis and future directions

The comprehensive research program documented in this thesis demonstrates the synergistic value of integrating clinical protocols, educational methodologies, and computational innovations to address fundamental challenges in critical care diagnostics. The consistent themes across all studies include the critical importance of standardization, rigorous validation against expert performance, clinical interpretability of technological solutions, and practical implementation considerations that account for real-world clinical workflows.

5.6.1 Clinical translation and implementation pathways

The immediate clinical translation pathway for this work involves the deployment of the AI ensemble model as a decision-support system in critical care settings. The perfect sensitivity and specificity demonstrated in validation, combined with the explainability features that align with clinical reasoning patterns, provide a strong foundation for clinical adoption. The BLUE-LUSS protocol offers an immediately implementable solution for quantitative lung monitoring at mechanically ventilated patients, with particular value in resource-constrained environments where comprehensive 12-zone protocols may be impractical. The cadaveric model provides a validated training platform that can accelerate competency development in lung POCUS, addressing the significant training gap that often limits ultrasound implementation. Objective 5 extends this pathway by showing how these validated components can be assembled into a compact clinician-

facing bedside system with local display, automation support, structured review, and traceable export.

5.6.2 Limitations and methodological considerations

While this research program has demonstrated robust outcomes across multiple domains, several limitations warrant consideration. Most importantly, the pneumothorax-positive class of the AI study comprised predominantly CT-verified cadaveric preparations and only two living patients with spontaneous pneumothorax; the possibility that the classifier partly exploits cadaver-versus-living differences rather than pneumothorax-specific features therefore cannot be excluded on statistical grounds alone. This concern is mitigated by two findings — the correct rejection of living breath-hold mimics that reproduce absent lung sliding with preserved lung pulse, and the pleural-line–localized Grad-CAM++ attention — but with so few living positives, detection of pneumothorax in living patients is demonstrated rather than statistically established. Prospective, multicenter validation in living patients is required before clinical deployment. The single-center nature of the clinical studies, though mitigated by the multi-source data approach, suggests the need for multicenter validation to confirm generalizability across different patient populations and healthcare systems. The expert panel, while substantially larger and more experienced than typically employed in similar studies, still represents a finite sample of clinical expertise. The current AI implementation focuses exclusively on M-mode analysis, and while this has proven sufficient for pneumothorax detection, integration with B-mode video analysis could enhance performance for other diagnostic applications. The cadaveric model, despite its high fidelity, cannot fully replicate the dynamic physiological conditions of living patients, particularly regarding cardiogenic oscillations and tissue perfusion characteristics.

5.6.3 Future research directions

This research establishes several promising directions for future investigation. The AI framework developed for pneumothorax detection should be expanded to encompass the full differential diagnosis of respiratory failure, including pulmonary edema, pneumonia, and pleural effusions. The ensemble approach could be adapted for continuous monitoring applications, potentially detecting developing pathologies before they become clinically apparent. The BLUE-LUSS protocol warrants validation in broader

critical care populations beyond COVID-19, including patients with ARDS from other causes and post-operative respiratory complications. The cadaveric model methodology could be extended to simulate other critical pathologies such as pleural effusions, consolidations, and pulmonary edema, creating a comprehensive library of simulated pathologies for training and AI development.

A further important direction is expansion of the underlying database through multicenter acquisition, broader device and probe diversity, and increased representation of technically difficult and borderline cases. Such expansion would not only support further optimization of model performance, but would also strengthen generalizability, regulatory preparedness, and future clinical implementation.

Additional architectural comparisons should also be included in future work. In particular, Vision Transformers and hybrid CNN-transformer models represent relevant candidates for evaluation once larger and more diverse datasets become available. Such comparisons should assess not only classification performance, but also explainability, computational demand, and suitability for compact edge deployment.

At the time of preparation of the defense version of this dissertation, the translational pathway had already progressed beyond a purely conceptual stage. A substantial development grant had been awarded within the Óbuda University Initium Nexus proof-of-concept program, with the current funding cycle running until January 31, 2027. This phase is expected to support the maturation of the system toward a regulation-aware implementation pathway and may accelerate future manufacturer adoption.

5.7 Summary of discussion

This comprehensive discussion has synthesized the findings from an integrated research program that bridges bedside clinical protocols with advanced bioengineering approaches. The development and expert benchmarking of an explainable AI ensemble model for pneumothorax detection establishes a new paradigm for medical AI systems, demonstrating that perfect diagnostic performance can be achieved while maintaining clinical interpretability (114). The validation of the BLUE-LUSS protocol provides a practical solution for quantitative lung monitoring in critically ill patients, balancing comprehensive assessment with clinical feasibility. The creation of a high-fidelity cadaveric model addresses fundamental challenges in ultrasound education and AI data generation. The proposed comprehensive translational framework and bedside runtime

architecture developed under Objective 5 complete this trajectory by showing how validated models, standardized acquisition, automation, and clinician review can be integrated into one deployable decision-support pathway. Collectively, these contributions represent scientific advancements in the field of critical care ultrasonography, providing both immediate clinical tools and foundational frameworks for future innovation. The successful integration of these diverse methodologies demonstrates the powerful synergy that can be achieved when clinical expertise, educational methodology, and computational innovation are unified toward common goals of improved diagnostic accuracy, enhanced patients' safety, and expanded access to expert-level care.

6. Conclusions

This doctoral thesis has systematically charted the integration of lung-focused POCUS with translational bioengineering, establishing a robust framework that bridges established clinical protocols with computational innovation. The research trajectory, documented across multiple peer-reviewed publications, demonstrates a coherent progression from foundational model development to clinical validation and, ultimately, to the creation and benchmarking of sophisticated AI decision-support systems. The collective findings address the persistent challenges in critical care diagnostics—namely, operator dependency, interpretive variability, and the exigent need for rapid, accurate decision-making at the bedside.

The foundational pillar of this work was the meticulous development and validation of a novel human cadaveric model for pneumothorax simulation. This model was engineered to overcome the significant ethical and practical limitations of training at critically ill patients or relying solely on less realistic phantoms. By utilizing a minimally invasive technique involving controlled intrapleural air insufflation via a small-bore cannula, the model successfully replicated the key sonographic signs of pneumothorax—specifically, the absence of lung sliding, the presence of A-lines, and the pathognomonic lung point. Expert evaluations by intensive care specialists with substantial ultrasound experience confirmed the model's high fidelity. Quantitative analysis using Cohen's κ revealed moderate to excellent inter-rater agreement across key metrics, including image quality, suitability for educational use, clarity of lung profiles, and feasibility for semi-quantitative scoring. This model has proven to be an invaluable asset, not only for structured skill acquisition in graduate and post-graduate medical education but also for generating a high-fidelity, standardized image database essential for subsequent computational research.

Building upon this anatomical and educational foundation, the thesis extensively detailed the clinical application and standardization of lung ultrasound within the high-stakes environments of critical care and perioperative medicine. A significant contribution was the proposal and the validation of the 'BLUE-LUSS' protocol, a streamlined, single-operator methodology designed for the unique constraints of the ICU. This protocol adapts the well-established BLUE points for semi-quantitative assessment using both the cLUSS and qLUSS lung ultrasound scores. In a prospective feasibility study on

mechanically ventilated COVID-19 patients, the BLUE-LUSS protocol demonstrated excellent inter-rater reliability, with ICCs reaching 0.92-0.93 for cLUSS and 0.79-0.87 for qLUSS. Furthermore, it exhibited significant inverse correlations with the PaO_2/FiO_2 ratio, a key indicator of oxygenation status, with the strongest correlation ($r = -0.55$, $p = 0.012$) observed for longitudinal qLUSS scans. This protocol effectively balances comprehensive lung assessment with practical feasibility, minimizing patient manipulation and staff exposure—a critical consideration during infectious disease outbreaks and in resource-limited settings.

The translational core of this thesis resides in the pioneering development, rigorous validation, and expert benchmarking of an explainable AI ensemble model for automated pneumothorax detection. Confronting the limitations of previous studies—such as small, non-diverse datasets and a lack of robust clinical comparison—this research leveraged a large, heterogeneous dataset of 1,856 M-mode images. This dataset was sourced from a wide spectrum of subjects, including critically ill ICU patients with various pathologies, healthy volunteers (providing normal and breath-hold "lung pulse" data), and the tailored human cadaveric models providing definitive pneumothorax-positive cases. The AI architecture employed a soft-voting ensemble of five convolutional neural networks (ResNet, Inception v3, Xception, Inception-ResNet-v2, VGG16), utilizing transfer learning from ImageNet and fine-tuning on the lung ultrasound domain.

The performance of this AI ensemble was excellent. On a meticulously balanced, independent test set, the model achieved perfect diagnostic metrics: 100% sensitivity (95% CI: 85.8%-100%) and 100% specificity (95% CI: 85.8%-100%). Then this performance was benchmarked against a panel of 11 highly experienced clinicians, each with over a decade of lung ultrasound practice. While the human experts demonstrated high aggregate sensitivity (95.8% in B-mode, 92.8% in M-mode), their specificity was significantly lower (81.0% in B-mode, 73.5% in M-mode). Statistical analysis using mixed-effects logistic regression confirmed that the AI's superiority in specificity was significant ($p=0.0194$), while its sensitivity was statistically non-inferior. No single expert could match the AI's combined perfect sensitivity and specificity; the best human experts were forced to trade one for the other. The AI effectively functioned as an idealized "expert committee," achieving consensus-level performance in every case and demonstrating a remarkable ability to correctly classify challenging scenarios, such as

subtle T-line signs from breath-holding that frequently triggered false positives among clinicians.

A critical finding was the analysis of diagnostic performance across imaging modes. Human experts performed significantly worse in specificity when using M-mode alone compared to B-mode ($p=0.00337$), highlighting a key area of interpretive vulnerability. However, the AI model, which was trained and evaluated exclusively on M-mode images, completely circumvented this limitation, proving that the information content within M-mode is sufficient for expert-level diagnosis when processed by a robust algorithm. Furthermore, the model demonstrated excellent generalizability, performing with equal proficiency on both cadaveric and live patients' scans, thereby validating the use of high-fidelity simulation models for AI training.

To bridge the gap between algorithmic "black boxes" and clinical trust, this research integrated explainable AI (XAI) methodologies. The application of Grad-CAM++ (Gradient-weighted Class Activation Mapping) provided visual explanations for the model's predictions, generating heatmaps that highlighted the image regions most influential in their decision. When five clinical experts evaluated these heatmaps, they demonstrated substantial inter-rater agreement (Fleiss' $\kappa = 0.750$) and rated 84.2% of the AI's focal areas as clinically relevant, meaning the AI was "looking at" the same sonographic landmarks—such as the pleural line and the regions below it characterizes the seashore, barcode, or T-line signs—that clinicians use for diagnosis. This transparency is paramount for fostering clinician acceptance and integrating AI as a trustworthy adjunct in the diagnostic process.

In a culminating translational step, the research has progressed beyond retrospective analysis to the development of a real-time, deployable bedside AI tool. This system, demonstrated in supplementary materials, showcases the practical integration of the ensemble model into a clinical workflow, providing immediate diagnostic support during ultrasound examinations.

This final translational stage also fulfilled the thesis objective of establishing a standardized implementation framework for AI-assisted lung POCUS. The bedside runtime unified disciplined acquisition settings, local edge hardware, touchscreen interaction, optional automation, clinician feedback, and traceable export into a single

decision-support pathway. As a result, this thesis did not end at model validation but progressed to a concrete clinician-in-the-loop deployment framework.

In summary, my thesis provides compelling evidence that the synergistic union of lung POCUS and AI-driven bioengineering can profoundly enhance diagnostic precision, support standardized clinical practice, and improve patient safety in critical care. It delivers a multi-faceted contribution: a validated educational model for skill development, a standardized and feasible clinical protocol for routine monitoring, and a benchmarked, explainable AI system that can augment human expertise. By systematically addressing the continuum from basic model development to clinical implementation and computational innovation, this work establishes a foundational framework for the future of intelligent, assistive technologies in bedside medicine. The methodologies and tools developed herein pave the way for wider adoption, promising to make expert-level diagnostic capability more accessible, consistent, and integrated into the dynamic and demanding environment of critical care, ultimately contributing to more efficient, equitable, and high-quality patient outcomes.

7. Summary

This thesis integrates clinical lung ultrasonography with translational bioengineering to address diagnostic challenges in intensive care medicine. It establishes a complete pipeline from anatomical model development, through clinical protocol validation, to a deployable AI system for bedside pneumothorax detection and lung assessment.

The 'BLUE-LUSS' protocol was introduced and validated as a streamlined, single-operator approach to semi-quantitative lung assessment in mechanically ventilated patients. It demonstrated excellent inter-rater reliability (ICC 0.79–0.93) and a significant inverse correlation with the PaO_2/FiO_2 ratio, while remaining feasible under the constraints of routine critical care.

A novel human cadaveric model was developed for pneumothorax simulation, using minimally invasive intrapleural air insufflation to reproduce realistic sonographic signs while preserving thoracic integrity. Expert evaluation confirmed its fidelity and educational utility, and the clip library serves both as training material and as standardized data for AI development.

The core bioengineering contribution was an explainable AI ensemble for pneumothorax detection. Trained on 1,856 M-mode images from ICU patients, healthy volunteers and cadaveric models, the soft-voting ensemble of five convolutional neural networks achieved 100% sensitivity and specificity on independent testing, and outperformed 11 expert clinicians in specificity with non-inferior sensitivity. Grad-CAM++ visualization made these decisions interpretable, with 84.2% of focal points judged clinically relevant by experts.

The work culminated in a deployable bedside system in which the trained models run on compact edge hardware under clinician review, with structured session export. Disciplined acquisition, workflow design, automation and traceable documentation define a practical decision-support pathway rather than a standalone algorithm. The thesis therefore describes not a device but an intelligent stethoscope: a bedside pathway in which standardized acquisition, realistic training material, explainable inference and clinician oversight operate together to improve diagnostic accuracy and patient safety in critical care.

8. References

1. Lau YH, See KC. Point-of-care ultrasound for critically-ill patients: A mini-review of key diagnostic features and protocols. World J Crit Care Med. 2022;11(2):70–84.
2. Lichtenstein DA. Lung Ultrasound in the Critically Ill. Springer International Publishing Switzerland 2016. 2016:XXXVI, 376.
3. Denault A, Girard M, Cavayas YA. Lung Ultrasound in the Critically Ill–The BLUE Protocol: Daniel A. Lichtenstein. Springer International Publishing Switzerland 2016; Hardcover 149eBook 109; 376 pages ISBN 978-3-319-15370-4. Springer; 2016.
4. Lichtenstein DA, Meziere GA. Relevance of lung ultrasound in the diagnosis of acute respiratory failure*: the BLUE protocol. Chest. 2008;134(1):117–25.
5. Lichtenstein DA. BLUE-protocol and FALLS-protocol: two applications of lung ultrasound in the critically ill. Chest. 2015;147(6):1659–70.
6. Lichtenstein DA. Lung Ultrasound in ARDS: The Pink-Protocol. The Place of Some Other Applications in the Intensive Care Unit (CLOT-Protocol, Fever-Protocol). Lung Ultrasound in the Critically Ill: The BLUE Protocol. Cham: Springer International Publishing; 2016. p. 203–16.
7. Lichtenstein DA. Lung Ultrasound in the Critically Ill The BLUE Protocol. 1 ed: Springer Cham; 2016. XXXVI, 376 p.
8. Lichtenstein DA, Malbrain M. Lung ultrasound in the critically ill (LUCI): A translational discipline. Anaesthesiol Intensive Ther. 2017;49(5):430–6.
9. Lichtenstein DA. Current Misconceptions in Lung Ultrasound: A Short Guide for Experts. Chest. 2019;156(1):21–5.
10. Mojoli F, Bouhemad B, Mongodi S, Lichtenstein D. Lung Ultrasound for Critically Ill Patients. Am J Respir Crit Care Med. 2019;199(6):701–14.
11. Mayo P, Copetti R, Feller-Kopman D, Mathis G, Maury E, Mongodi S, et al. Thoracic ultrasonography: a narrative review. Intensive care medicine. 2019;45:1200–11.
12. Łyźniak P, Świętoń D, Serafin Z, Szurowska E. Lung ultrasound in a nutshell. Lines, signs, some applications, and misconceptions from a radiologist's point of view. Polish journal of radiology. 2023;88:e294.
13. Alrajab S, Youssef AM, Akkus NI, Caldito G. Pleural ultrasonography versus chest radiography for the diagnosis of pneumothorax: review of the literature and meta-analysis. Critical Care. 2013;17(5):R208.
14. Blaivas M, Lyon M, Duggal S. A prospective comparison of supine chest radiography and bedside ultrasound for the diagnosis of traumatic pneumothorax. Academic Emergency Medicine. 2005;12(9):844–9.
15. Chen L, Zhang Z. Bedside ultrasonography for diagnosis of pneumothorax. Quantitative imaging in medicine and surgery. 2015;5(4):618.
16. Husain LF, Hagopian L, Wayman D, Baker WE, Carmody KA. Sonographic diagnosis of pneumothorax. J Emerg Trauma Shock. 2012;5(1):76–81.
17. Volpicelli G. Sonographic diagnosis of pneumothorax. Intensive care medicine. 2011;37:224–32.
18. Volpicelli G, Elbarbary M, Blaivas M, Lichtenstein DA, Mathis G, Kirkpatrick AW, et al. International evidence-based recommendations for point-of-care lung ultrasound. Intensive care medicine. 2012;38:577–91.
19. Vetrugno L, Mojoli F, Boero E, Berchialla P, Bignami EG, Orso D, et al. Level of Diffusion and Training of Lung Ultrasound during the COVID-19 Pandemic - A National

Online Italian Survey (ITALUS) from the Lung Ultrasound Working Group of the Italian Society of Anesthesia, Analgesia, Resuscitation, and Intensive Care (SIAARTI). Ultraschall Med. 2021.
20. Gargani L, Soliman-Aboumarie H, Volpicelli G, Corradi F, Pastore MC, Cameli M. Why, when, and how to use lung ultrasound during the COVID-19 pandemic: enthusiasm and caution. Eur Heart J Cardiovasc Imaging. 2020;21(9):941–8.
21. Hosseini-Nik H, Bayanati H, Souza CA, Gupta A, McInnes MDF, Pena E, et al. Limited Chest Ultrasound to Replace CXR in Diagnosis of Pneumothorax Post Image-Guided Transthoracic Interventions. Canadian Association of Radiologists Journal. 2022;73(2):403–9.
22. Yousuf B, Sujatha KS, Alfoudri H, Mansurov V. Transport of critically ill COVID-19 patients. Intensive Care Med. 2020;46(8):1663–4.
23. Alcantara MLd, Bernardo MPL, Autran TB, Lustosa RdP, Tayah M, Chagas LA, et al. Lung Ultrasound as a Triage Tool in an Emergency Setting during the Covid-19 Outbreak: comparison with CT Findings. International Journal of Cardiovascular Sciences. 2020;33(5):479–87.
24. Boccatonda A, Grignaschi A, Lanotte AMG, Cocco G, Vidili G, Giostra F, et al. Role of Lung Ultrasound in the Management of Patients with Suspected SARS-CoV-2 Infection in the Emergency Department. J Clin Med. 2022;11(8).
25. Bosso G, Allegorico E, Pagano A, Porta G, Serra C, Minerva V, et al. Lung ultrasound as diagnostic tool for SARS-CoV-2 infection. Intern Emerg Med. 2021;16(2):471–6.
26. Chan KK, Joo DA, McRae AD, Takwoingi Y, Premji ZA, Lang E, et al. Chest ultrasonography versus supine chest radiography for diagnosis of pneumothorax in trauma patients in the emergency department. Cochrane Database Syst Rev. 2020;7(7):Cd013031.
27. Dargent A, Chatelain E, Si-Mohamed S, Simon M, Baudry T, Kreitmann L, et al. Lung ultrasound score as a tool to monitor disease progression and detect ventilator-associated pneumonia during COVID-19-associated ARDS. Heart Lung. 2021;50(5):700–5.
28. Ji L, Cao C, Gao Y, Zhang W, Xie Y, Duan Y, et al. Prognostic value of bedside lung ultrasound score in patients with COVID-19. Crit Care. 2020;24(1):700.
29. Peixoto AO, Costa RM, Uzun R, Fraga AMA, Ribeiro JD, Marson FAL. Applicability of lung ultrasound in COVID-19 diagnosis and evaluation of the disease progression: A systematic review. Pulmonology. 2021;27(6):529–62.
30. Bellani G, Rouby J-J, Constantin J-M, Pesenti A. Looking closer at acute respiratory distress syndrome: the role of advanced imaging techniques. Current Opinion in Critical Care. 2017;23(1):30–7.
31. Bouhemad B, Mongodi S, Via G, Rouquette I. Ultrasound for "lung monitoring" of ventilated patients. Anesthesiology. 2015;122(2):437–47.
32. Monastesse A, Girard F, Massicotte N, Chartrand-Lefebvre C, Girard M. Lung Ultrasonography for the Assessment of Perioperative Atelectasis: A Pilot Feasibility Study. Anesth Analg. 2017;124(2):494–504.
33. Lichtenstein DA, Mezière G, Lascols N, Biderman P, Courret J-P, Gepner A, et al. Ultrasound diagnosis of occult pneumothorax. Critical care medicine. 2005;33(6):1231–8.
34. Lichtenstein DA. Lung Ultrasound (in the Critically Ill) Superior to CT: the Example of Lung Sliding. Korean J Crit Care Med. 2017;32(1):1–8.
35. Gargani L, Volpicelli G. How I do it: lung ultrasound. Cardiovasc Ultrasound. 2014;12:25.

36. Mongodi S, Bouhemad B, Orlando A, Stella A, Tavazzi G, Via G, et al. Modified Lung Ultrasound Score for Assessing and Monitoring Pulmonary Aeration. Ultraschall Med. 2017;38(5):530–7.
37. Mongodi S, Santangelo E, De Luca D, Rovida S, Corradi F, Volpicelli G, et al. Quantitative Lung Ultrasound: Time for a Consensus? Chest. 2020;158(2):469–70.
38. Mongodi S, De Luca D, Colombo A, Stella A, Santangelo E, Corradi F, et al. Quantitative Lung Ultrasound: Technical Aspects and Clinical Applications. Anesthesiology. 2021;134(6):949–65.
39. Brat R, Yousef N, Klifa R, Reynaud S, Shankar Aguilera S, De Luca D. Lung Ultrasonography Score to Evaluate Oxygenation and Surfactant Need in Neonates Treated With Continuous Positive Airway Pressure. JAMA Pediatr. 2015;169(8):e151797.
40. Chiumello D, Mongodi S, Algieri I, Vergani GL, Orlando A, Via G, et al. Assessment of Lung Aeration and Recruitment by CT Scan and Ultrasound in Acute Respiratory Distress Syndrome Patients*. Critical Care Medicine. 2018;46(11):1761–8.
41. Costamagna A, Pivetta E, Goffi A, Steinberg I, Arina P, Mazzeo AT, et al. Clinical performance of lung ultrasound in predicting ARDS morphology. Ann Intensive Care. 2021;11(1):51.
42. Dargent A, Chatelain E, Kreitmann L, Quenot JP, Cour M, Argaud L. Lung ultrasound score to monitor COVID-19 pneumonia progression in patients with ARDS. PLoS One. 2020;15(7):e0236312.
43. Boero E, Gargani L, Schreiber A, Rovida S, Martinelli G, Maggiore SM, et al. Lung ultrasound among Expert operator'S: ScOring and iNter-rater reliability analysis (LESSON study) a secondary COWS study analysis from ITALUS group. Journal of Anesthesia, Analgesia and Critical Care. 2024;4(1):50.
44. Volpicelli G, Gargani L, Zieleskiewicz L, Tuinman PR, Kimura BJ, Zisis G, et al. International evidence-based recommendations for point-of-care lung ultrasound. Intensive Care Medicine. 2026;52(7):1415–46.
45. Demi L, Wolfram F, Klersy C, De Silvestri A, Ferretti VV, Muller M, et al. New international guidelines and consensus on the use of lung ultrasound. Journal of Ultrasound in Medicine. 2023;42(2):309–44.
46. Bruno A, Ignesti G, Salvetti O, Moroni D, Martinelli M. Efficient lung ultrasound classification. Bioengineering. 2023;10(5):555.
47. Cheng D, Lam EY. Transfer learning U-Net deep learning for lung ultrasound segmentation. arXiv preprint arXiv:211002196. 2021.
48. Yan WQ. Transfer Learning and Ensemble Learning. Computational Methods for Deep Learning: Theory, Algorithms, and Implementations. Singapore: Springer Nature Singapore; 2023. p. 191–203.
49. Yi Z, Wang Y, editors. Transfer learning on interstitial lung disease classification. 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML); 2021: IEEE.
50. Akinbo RS, Daramola OA. Ensemble machine learning algorithms for prediction and classification of medical images. Algorithms, Models and Applications. 2021:59.
51. Arnaout R, Curran L, Zhao Y, Levine JC, Chinn E, Moon-Grady AJ. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nat Med. 2021;27(5):882–91.
52. Manconi A, Armano G, Gnocchi M, Milanesi L. A soft-voting ensemble classifier for detecting patients affected by COVID-19. Applied Sciences. 2022;12(15):7554.

53. Borys K, Schmitt YA, Nauta M, Seifert C, Krämer N, Friedrich CM, et al. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. European journal of radiology. 2023;162:110786.
54. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors. 2023;23(2):634.
55. Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis. 2022;79:102470.
56. Brusasco C, Santori G, Bruzzo E, Trò R, Robba C, Tavazzi G, et al. Quantitative lung ultrasonography: a putative new algorithm for automatic detection and quantification of B-lines. Crit Care. 2019;23(1):288.
57. Ranaweera M, Mahmoud QH. Virtual to real-world transfer learning: A systematic review. Electronics. 2021;10(12):1491.
58. Orosz G, Szabó RZ, Szabó M, Gyombolai P, Tóth JT, Ruttkay T, et al. Explainable transfer learning ensemble AI model for lung ultrasound pneumothorax detection with expert benchmark. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2026.
59. Griffiths E. Helicopter emergency medical services use of thoracic point of care ultrasound for pneumothorax: a systematic review and meta-analysis. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine. 2021;29(1):163.
60. Jaščur M, Bundzel M, Malík M, Dzian A, Ferenčík N, Babič F. Detecting the absence of lung sliding in lung ultrasounds using deep learning. Applied Sciences. 2021;11(15):6976.
61. Avila J, Smith B, Mead T, Jurma D, Dawson M, Mallin M, et al. Does the Addition of M-Mode to B-Mode Ultrasound Increase the Accuracy of Identification of Lung Sliding in Traumatic Pneumothoraces? Journal of Ultrasound in Medicine. 2018;37(11):2681–7.
62. Berlet T, Etter R. Favourable Experience with M-Mode Sonography in the Diagnosis of Pneumothorax in Two Patients with Thoracic Subcutaneous Emphysema. Case reports in radiology. 2014;2014(1):906127.
63. Sugibayashi T, Walston SL, Matsumoto T, Mitsuyama Y, Miki Y, Ueda D. Deep learning for pneumothorax diagnosis: a systematic review and meta-analysis. European Respiratory Review. 2023;32(168).
64. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.
65. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in medicine. 2002;21(16):2409–19.
66. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. American journal of epidemiology. 2018;187(4):864–70.
67. 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. 2022;17(10):e0276213.
68. Gil-Rodríguez J, Pérez de Rojas J, Aranda-Laserna P, Benavente-Fernández A, Martos-Ruiz M, Peregrina-Rivas JA, et al. Ultrasound findings of lung ultrasonography in COVID-19: A systematic review. Eur J Radiol. 2022;148:110156.
69. Soldati G, Smargiassi A, Inchingolo R, Buonsenso D, Perrone T, Briganti DF, et al. Proposal for International Standardization of the Use of Lung Ultrasound for Patients

With COVID-19: A Simple, Quantitative, Reproducible Method. J Ultrasound Med. 2020;39(7):1413–9.
70. Soldati G, Smargiassi A, Inchingolo R, Buonsenso D, Perrone T, Briganti DF, et al. Time for a new international evidence-based recommendations for point-of-care lung ultrasound. J Ultrasound Med. 2021;40(2):433–4.
71. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002;21(9):1331–5.
72. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155–63.
73. Xue H, Li C, Cui L, Tian C, Li S, Wang Z, et al. M-BLUE protocol for coronavirus disease-19 (COVID-19) patients: interobserver variability and correlation with disease severity. Clin Radiol. 2021;76(5):379–83.
74. Szabó RZ, Orosz G, Ungi T, Barr C, Yeung C, Incze R, et al., editors. Automation of lung ultrasound imaging and image processing for bedside diagnostic examinations. 2023 IEEE 17th International Symposium on Applied Computational Intelligence and Informatics (SACI); 2023: IEEE.
75. Orosz G, Szabó RZ, Ungi T, Barr C, Yeung C, Fichtinger G, et al. Lung ultrasound imaging and image processing with artificial intelligence methods for bedside diagnostic examinations. Acta Polytech Hung. 2023;20(8):69–87.
76. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30(9):1323–41.
77. Ádám N, Orosz G, Berczi M, Ruttkay T. Proposal of a human cadaveric model for bedside ultrasound-based pneumothorax detection. Orvosi Hetilap. 2023;164(46):1824–30.
78. Szabó M, Orosz G, Iványi ZD, Darvas K. A mellkas és a tüdő ultrahangvizsgálatának szerepe az általános érzéstelenítés során. Orvosi Hetilap. 2023;164(22):864–70.
79. Hoyer R, Means R, Robertson J, Rappaport D, Schmier C, Jones T, et al. Ultrasound-guided procedures in medical education: a fresh look at cadavers. Intern Emerg Med. 2016;11(3):431–6.
80. Skulec R, Parizek T, David M, Matousek V, Cerny V. Lung Point Sign in Ultrasound Diagnostics of Pneumothorax: Imitations and Variants. Emerg Med Int. 2021;2021:6897946.
81. ter Haar G. The new British Medical Ultrasound Society Guidelines for the safe use of diagnostic ultrasound equipment. Ultrasound. 2010;18(2):50–1.
82. Smith III BC, Avila J. M. mode. ify: A Free Online Tool to Generate Post Hoc M-Mode Images From Any Ultrasound Clip. Journal of Ultrasound in Medicine. 2016;35(2):435–9.
83. Szegedy C, Ioffe S, Vanhoucke V, Alemi A, editors. Inception-v4, inception-resnet and the impact of residual connections on learning. Proceedings of the AAAI conference on artificial intelligence; 2017.
84. Chollet F, editor Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
85. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, editors. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
86. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

87. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014.
88. Rajaraman S, Zamzmi G, Antani SK. Novel loss functions for ensemble-based medical image classification. Plos one. 2021;16(12):e0261307.
89. Cardoso MJ, Li W, Brown R, Ma N, Kerfoot E, Wang Y, et al. Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:221102701. 2022.
90. Collett D. Modelling binary data: CRC press; 2002.
91. Coughlin SS, Trock B, Criqui MH, Pickle LW, Browner D, Tefft MC. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. Journal of clinical epidemiology. 1992;45(1):1–7.
92. Zhang T, Yan S, Wei G, Yang L, Yu T, Ma Y. Automatic diagnosis of pneumothorax with M-mode ultrasound images based on D-MPL. International Journal of Computer Assisted Radiology and Surgery. 2023;18(2):303–12.
93. VanBerlo B, Wu D, Li B, Rahman MA, Hogg G, VanBerlo B, et al. Accurate assessment of the lung sliding artefact on lung ultrasonography using a deep learning approach. Computers in biology and medicine. 2022;148:105953.
94. VanBerlo B, Li B, Wong A, Hoey J, Arntfield R, editors. Exploring the utility of self-supervised pretraining strategies for the detection of absent lung sliding in m-mode lung ultrasound. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023.
95. Kumar A, Weng Y, Graglia S, Chung S, Duanmu Y, Lalani F, et al. Interobserver Agreement of Lung Ultrasound Findings of COVID-19. J Ultrasound Med. 2021;40(11):2369–76.
96. Lerchbaumer MH, Lauryn JH, Bachmann U, Enghard P, Fischer T, Grune J, et al. Point-of-care lung ultrasound in COVID-19 patients: inter- and intra-observer agreement in a prospective observational study. Sci Rep. 2021;11(1):10678.
97. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN, editors. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE winter conference on applications of computer vision (WACV); 2018: IEEE.
98. Peng QY, Liu LX, Zhang Q, Zhu Y, Zhang HM, Yin WH, et al. Lung ultrasound score based on the BLUE-plus protocol is associated with the outcomes and oxygenation indices of intensive care unit patients. J Clin Ultrasound. 2021;49(7):704–14.
99. Heldeweg MLA, Lieveld AWE, de Grooth HJ, Heunks LMA, Tuinman PR. Determining the optimal number of lung ultrasound zones to monitor COVID-19 patients: can we keep it ultra-short and ultra-simple? Intensive Care Med. 2021;47(9):1041–3.
100. Perrone T, Soldati G, Padovini L, Fiengo A, Lettieri G, Sabatini U, et al. A New Lung Ultrasound Protocol Able to Predict Worsening in Patients Affected by Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia. Journal of Ultrasound in Medicine. 2021;40(8):1627–35.
101. Pierrakos C, Lieveld A, Pisani L, Smit MR, Heldeweg M, Hagens LA, et al. A Lower Global Lung Ultrasound Score Is Associated with Higher Likelihood of Successful Extubation in Invasively Ventilated COVID-19 Patients. Am J Trop Med Hyg. 2021;105(6):1490–7.
102. Pisani L, Vercesi V, van Tongeren PSI, Lagrand WK, Leopold SJ, Huson MAM, et al. The diagnostic accuracy for ARDS of global versus regional lung ultrasound scores - a post hoc analysis of an observational study in invasively ventilated ICU patients. Intensive Care Med Exp. 2019;7(Suppl 1):44.

103. Wangüemert Pérez AL, Figueira Gonçalves JM, Hernández Pérez JM, Ramallo Fariña Y, Del Castillo Rodriguez JC. Prognostic value of lung ultrasound and its link with inflammatory biomarkers in patients with SARS-CoV-2 infection. Respir Med Res. 2021;79:100809.
104. Eisenmann S, Winantea J, Karpf-Wissel R, Funke F, Stenzel E, Taube C, et al. Thoracic Ultrasound for Immediate Exclusion of Pneumothorax after Interventional Bronchoscopy. Journal of Clinical Medicine. 2020;9(5):1486.
105. Laursen CB, Pietersen PI, Jacobsen N, Falster C, Juul AD, Davidsen JR. Lung ultrasound assessment for pneumothorax following transbronchial lung cryobiopsy. ERJ Open Research. 2021;7(3):00045–2021.
106. Lindsey T, Lee R, Grisell R, Vega S, Veazey S, editors. Automated pneumothorax diagnosis using deep neural networks. Iberoamerican congress on pattern recognition; 2018: Springer.
107. Mento F, Perrone T, Macioce VN, Tursi F, Buonsenso D, Torri E, et al. On the Impact of Different Lung Ultrasound Imaging Protocols in the Evaluation of Patients Affected by Coronavirus Disease 2019: How Many Acquisitions Are Needed? J Ultrasound Med. 2021;40(10):2235–8.
108. Rouby J-J, Arbelot C, Gao Y, Zhang M, Lv J, An Y, et al. Training for lung ultrasound score measurement in critically ill patients. American Journal of Respiratory and Critical Care Medicine. 2018;198(3):398–401.
109. Fuchs L, Galante O, Almog Y, Dayan RR, Smoliakov A, Ullman Y, et al. Point of Care Lung Ultrasound Injury Score-A simple and reliable assessment tool in COVID-19 patients (PLIS I): A retrospective study. PLoS One. 2022;17(5):e0267506.
110. Székely R, Suhai F, Karlinger K, Baksa G, Szabaczki B, Bárány L, et al. Human Cadaveric Artificial Lung Tumor-Mimic Training Model. Pathology and Oncology Research. 2021;27.
111. Adhikari S, Zeger W, Wadman M, Walker R, Lomneth C. Assessment of a human cadaver model for training emergency medicine residents in the ultrasound diagnosis of pneumothorax. Biomed Res Int. 2014;2014:724050.
112. Auinger D, Orlob S, Wittig J, Honnef G, Heschl S, Feigl G, et al. Pneumothorax in a Thiel cadaver model of cardiopulmonary resuscitation. World J Emerg Med. 2023;14(2):143–7.
113. Lim D, Bartlett S, Horrocks P, Grant-Wakefield C, Kelly J, Tippett V. Enhancing paramedics procedural skills using a cadaveric model. BMC Medical Education. 2014;14(1):138.
114. Jin W, Li X, Fatehi M, Hamarneh G. Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical image analysis. 2023;84:102684.

9. Bibliography of the candidate's publications

9.1 Publications fundamentally connected to the thesiswork

Orosz, Gábor ✉ ; Szabó, Róbert Zsolt ; Szabó, Marcell ; Gyombolai, Pál ; Tóth, József T ; Ruttkay, Tamás ; Ferenci, Tamás ; Ungi, Tamás ; Fichtinger, Gábor ; Haidegger, Tamás
Explainable transfer learning ensemble AI model for lung ultrasound pneumothorax detection with expert benchmark
SCANDINAVIAN JOURNAL OF TRAUMA RESUSCITATION AND EMERGENCY MEDICINE Paper (2026)
DOI PubMed
Scopus - Emergency Medicine SJR: D1
Scopus - Critical Care and Intensive Care Medicine SJR: Q1

Ádám, N. ; Orosz, G. ; Berczi, M. ; Ruttkay, T. ✉
Humán kadávermodell a légmell ágy melletti ultrahang-diagnosztikájához: Előtanulmány
ORVOSI HETILAP 164 : 46 pp. 1824-1830. , 7 p. (2023)
DOI WoS REAL Scopus PubMed
Scopus - Medicine (miscellaneous) SJR: Q4

Orosz, Gábor ✉ ; Szabó, Róbert Zsolt* ; Ungi, Tamás ; Barr, Colton ; Yeung, Chris ; Fichtinger, Gábor ; Gál, János ; Haidegger, Tamás
Lung Ultrasound Imaging and Image Processing with Artificial Intelligence Methods for Bedside Diagnostic Examinations
ACTA POLYTECHNICA HUNGARICA 20 : 8 pp. 69-87. , 19 p. (2023)
DOI WoS Scopus
Scopus - Engineering (miscellaneous) SJR: Q2
Scopus - Multidisciplinary SJR: Q2

Szabó, Marcell ✉ ; Orosz, Gábor ; Iványi, Zsolt Dániel ; Darvas, Katalin
A mellkas és a tüdő ultrahangvizsgálatának szerepe az általános érzéstelenítés során
ORVOSI HETILAP 164 : 22 pp. 864-870. , 7 p. (2023)
DOI WoS REAL Scopus PubMed
Tudományos
Scopus - Medicine (miscellaneous) SJR: Q4

Szabó, Róbert Zsolt ✉ ; Orosz, Gábor ; Ungi, Tamás ; Barr, Colton ; Yeung, Chris ; Incze, Roland ; Fichtinger, Gabor ; Gál, János ; Haidegger, Tamás
Automation of lung ultrasound imaging and image processing for bedside diagnostic examinations
In: Szakál, Anikó (szerk.) IEEE 17th International Symposium on Applied Computational Intelligence and Informatics SACI 2023 : Proceedings
Budapest, Magyarország : Óbudai Egyetem, IEEE Hungary Section (2023) 818 p. pp. 779-783. , 6 p.
DOI Scopus

Orosz, Gábor ✉ ; Gyombolai, Pál ; Tóth, József T ; Szabó, Marcell
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 Paper: e0276213 , 15 p. (2022)
DOI WoS Scopus PubMed
Scopus - Multidisciplinary SJR: Q1

∑IF publications related to the thesis: 10.3.

9.2 Publications independent from the thesiswork

Bódis, Fruzsina ✉ ; Orosz, Gábor ; Tóth, József T. ; Szabó, Marcell ; Élő, László Gergely ; Gál, János ; Élő, Gábor
Percutaneous tracheostomy: Comparison of three different methods with respect to tracheal cartilage injury in cadavers—Randomized controlled study
PATHOLOGY AND ONCOLOGY RESEARCH 29 Paper: 1610934 , 8 p. (2023)
DOI WoS Scopus PubMed
Scopus - Medicine (miscellaneous) SJR: Q2
Scopus - Oncology SJR: Q2
Scopus - Pathology and Forensic Medicine SJR: Q2
Scopus - Cancer Research SJR: Q3

Bódis, Fruzsina ✉ ; Orosz, Gábor ; Szabó, Marcell ; Molnár, Viktória ; Tóth, T. József ; Élő, László Gergely ; Tamás, László ; Élő, Gábor
Az antropometriai paraméterek szerepe a cadaveren végzett percutan tracheostomás módszerekben [Role of anthropometric parameters in percutaneous tracheostomies performed on cadavers]
ORVOSI HETILAP 164 : 16 pp. 630-635. , 6 p. (2023)
DOI WoS REAL Scopus PubMed Matarka
Scopus - Medicine (miscellaneous) SJR: Q4

∑IF publications not related to the thesis: 3.1.

10. Acknowledgements

I would first like to express my deepest gratitude to *my Family*. I am especially grateful to my wife, *Kinga*, and to our children, *Johanna* and *Marcell*, for their patience, love, and constant motivation throughout the years of this work. Their presence gave meaning and perspective to every milestone of the dissertation.

I am deeply thankful to *my Father* for the many hours of childcare that made sustained work on this dissertation possible. His practical help provided the time and stability without which much of this work could not have been completed. I also remember *my Mother*, who could no longer be by my side during this journey, but I am certain that she is now watching me with pride from above.

I owe special thanks to my supervisor, *Tamás Haidegger*, for recognizing the value of this research direction from the very beginning, for supporting my work with every available means, and for standing beside me not only as a mentor but also as a friend.

I am sincerely grateful to my research engineer colleague, *Róbert Zsolt Szabó*, whose engineering insight and readiness to help were indispensable whenever my ideas had to be transformed into practical technical solutions.

I am also deeply grateful to *Gábor Fichtinger* and *Tamás Ungi* for welcoming me on two occasions at the Queen's University School of Computing and for providing outstanding support that greatly contributed to my ability to adopt an engineering way of thinking alongside my clinical perspective.

I would also like to thank my former and current supervisors for their continued support, understanding, and trust during the different phases of this doctoral work.

My appreciation also goes to our small research team, *Pál Gyombolai*, *József T. Tóth*, and *Marcell Szabó*, for their perseverance, commitment, and shared effort throughout the demanding phases of the project.

I am grateful to *Tamás Ruttkay* and *Eszter Hatfaludy* for their availability, collaboration, and support in the cadaver-based research activities, which were essential to one of the most distinctive branches of this dissertation.

Finally, I would like to thank all *my friends and colleagues* who cannot be listed here individually, yet whose support, encouragement, and contributions have played an important part in the success of my research.

Appendix

Table 1. Thesis objectives, datasets, methods, and outputs.

Objective	Clinical / Engineering question	Dataset / Study population	Main method	Main outcome / output	Associated sections
Objective 1	Can an AI ensemble outperform human experts in PTX detection?	1,856 M-mode images from 114 subjects (ICU patients, volunteers, cadavers)	Soft-voting ensemble of five CNNs with Grad-CAM++ explainability	100% sensitivity and specificity; significant superiority in specificity over 11 experts	2.1, 3.3.1-3.3.3, 4.1, 5.1
Objective 2	Is the BLUE-LUSS protocol reliable and feasible for mechanically ventilated COVID-19 patients?	24 COVID-19 ICU patients enrolled, 20 analyzed; 400 ultrasound loops	Single-operator, three-point-per-hemithorax LUS using cLUSS and qLUSS scoring	Excellent reliability (ICC 0.79-0.93); significant correlation with PaO2/FiO2 ratio	2.2, 3.1.1, 4.2, 5.2
Objective 3	Can AI automate LUS image processing and be deployed at the bedside?	1,097 manually segmented frames from 40 patients	U-Net semantic segmentation; Raspberry Pi 5 edge-runtime	87.53% segmentation accuracy; deployable real-time bedside prototype	2.3, 3.3.4-3.3.5, 4.3, 5.3

Objective	Clinical / Engineering question	Dataset / Study population	Main method	Main outcome / output	Associated sections
			with HAILO pathway		
Objective 4	Can a human cadaveric model provide realistic PTX simulation for training and AI development?	5 fresh unembalmed cadavers; 141 ultrasound clips	Minimally invasive pleural cannulation with controlled air insufflation	High clinical fidelity; substantial expert agreement (kappa = 0.763) for training utility	2.4, 3.1.2, 4.4, 5.4
Objective 5	How can AI and bioengineering be translated into a standardized bedside lung POCUS framework?	Cross-objective synthesis of the clinical, cadaveric, AI, and bedside runtime branches	Standardized acquisition framework; Pi-local edge-runtime with AV.io capture, touchscreen UI, HAILO acceleration, automation,	Deployable clinician-in-the-loop bedside decision-support framework with manual and automation-enabled workflows	2.5, 3.2.1-3.3.5, 4.5, 5.5

Objective	Clinical / Engineering question	Dataset / Study population	Main method	Main outcome / output	Associated sections
			and provenance layer		

Table 2. Study populations and data assets.

Cohort / Asset	N (subjects)	Time period	Data type / Count	Role in thesis	Linked objective(s)
ICU clinical AI cohort	114 total subjects	2020-2024	1,856 M-mode images (post-exclusion)	Core training and testing data for AI classification	Objective 1
COVID-19 BLUE-LUSS cohort	20 patients analyzed (24 enrolled)	Dec 2021 - Apr 2022	400 B-mode ultrasound loops	Protocol feasibility, inter-rater reliability, and clinical correlation	Objective 2
Healthy volunteers	10 physicians	N/A	112 M-mode images (lung pulse / breath-hold)	Baseline data and challenging case generation (T-lines)	Objective 1
Cadaveric model cohort	5 fresh cadavers	2021-2022	141 evaluation clips / 826	Simulation validation and AI ground-	Objectives 1 and 4

Cohort / Asset	N (subjects)	Time period	Data type / Count	Role in thesis	Linked objective(s)
			AI M-mode images	truth generation	
Segmentation dataset	40 patients (22 COVID, 18 non-COVID)	2020-2024	1,097 manually segmented frames	U-Net training and leave-one-out cross-validation	Objective 3
Expert benchmark set	48 cases (24 PTX, 24 no PTX)	N/A	48 B-mode clips plus 48 M-mode images	Independent benchmarking of AI versus 11 expert clinicians	Objective 1

Table 3. Ultrasound systems and standardized acquisition settings.

Ultrasound system	Transducer	Frequency range	Exam use	Key standardized settings	Disabled post-processing features
Philips CX50	C5-1 convex	1.0-5.0 MHz	BLUE-LUSS and AI clinical cohort	Depth: pleural line + 3-8 cm; single focus at pleural line; gain optimized for primary artifacts; MI < 0.7; TIs < 1.0	THI, XRes, CrossXBeam, and SRI disabled

Ultrasound system	Transducer	Frequency range	Exam use	Key standardized settings	Disabled post-processing features
GE Venue GO	C1-5-RS convex	1.4-5.7 MHz	BLUE-LUSS and cadaveric studies	Depth: pleural line + 3-8 cm; single focus at pleural line; gain optimized for primary artifacts; MI < 0.7; TIs < 1.0	THI, XRes, CrossXBeam, and SRI disabled

Table 4. AI models, training strategy, and evaluation design.

Component	Architecture / Method	Input	Output	Split / Validation strategy	Key hyperparameters	Evaluation endpoint
Ensemble classifier	Soft-voting ensemble (ResNet50, Inception v3, Xception, Inception-ResNet-v2, VGG16)	256 x 256 M-mode image	Four-class probability distribution	80/20 subject-level split; five-fold cross-validation during training	Batch size 32; initial learning rate 1e-4 with cosine annealing; Adam optimizer	Sensitivity, specificity, and accuracy versus 11 experts

Component	Architecture / Method	Input	Output	Split / Validation strategy	Key hyperparameters	Evaluation endpoint
Explainability	Grad-CAM++ with ensemble-level map averaging	256 x 256 M-mode image	Coarse localization heatmap	Evaluated on 48 independent benchmark cases	Activation maps averaged across the five ensemble members	Expert-rated clinical relevance and Fleiss' kappa
Segmentation	U-Net encoder-decoder with skip connections and zero-padding	256 x 256 B-mode frame	Six semantic-class masks	Leave-one-out cross-validation at patient level	Batch size 8; learning rate 1e-4; 16 base filters; augmentation enabled	Validation accuracy and class-wise Dice performance
Bedside runtime	Raspberry Pi 5 edge runtime with AV.io capture and optional HAILO HEF branch	Live ultrasound stream	Real-time localization, M-mode generation, prediction, and optional Grad-CAM++	Iterative hardware-in-the-loop deployment and bedside validation	YOLO ROI detection; four-model compiled classifier; manual and automation modes	Functional bedside operation and structured provenance export

Table 6. Objective 2 baseline cohort characteristics.

Characteristics of the mechanically ventilated COVID-19 cohort used for BLUE-LUSS validation.

Variable	Value / Median (IQR)	Count (%)
Patients enrolled / analyzed	24 / 20	N/A
Total ultrasound loops recorded	400	N/A
Age (years)	60 (46.5-67.5)	N/A
Sex (male)	N/A	15 (75.0%)
Vaccination status (none)	N/A	14 (70.0%)
CT lung involvement (>50%)	N/A	10 (50.0%)
APACHE II score	11.25 (6-32)	N/A
CURB-65 score	2 (0-4)	N/A
WBC at admission (G/l)	9.4 (6.5-12.5)	N/A
CRP at admission (mg/l)	163.2 (69.8-234.5)	N/A

Table 8. Objective 3 segmentation and bedside deployment summary.

Development phase / Branch	Core technology and hardware	Input / Dataset	Key functions and validation	Main outcomes and performance
1. Semantic segmentation (algorithm training)	U-Net deep neural network with zero-padding and encoder-decoder structure	1,097 manually segmented B-mode frames from 40 patients	Leave-one-out cross-validation; pixel-wise segmentation of BLUE-relevant signs	87.53% validation accuracy; excellent performance for ribs and pleura, moderate performance for early B-lines

Development phase / Branch	Core technology and hardware	Input / Dataset	Key functions and validation	Main outcomes and performance
2. Manual Pi-local baseline (first prototype)	Raspberry Pi 5, AV.io capture, local touchscreen	Live ultrasound stream captured from the ultrasound machine	Manual ROI cropping, YOLO localization, manual M-mode generation	Functional edge runtime established; CPU-based Grad-CAM++ generated on demand
3. HAILO-accelerated app (hardware acceleration)	Raspberry Pi 5 with HAILO AI accelerator	Live ultrasound stream	HEF-compiled YOLO detection and four-model ensemble prediction	Fast bedside inference; successful hybrid architecture combining edge AI and host-side explainability
4. Automation-enabled app (final bedside tool)	Raspberry Pi 5 with touchscreen dashboard	Live ultrasound stream	Pleural tracking, consensus scanline derivation, automated M-mode capture, structured review workflow	High automation with fail-safes; captures user feedback and provenance metadata

Table 10. Integrated synthesis across objectives.

Objective	Unmet need	Innovation	Main quantitative result	Translational relevance
Objective 1 (AI model)	Operator dependency and false positives in PTX detection	Soft-voting ensemble with Grad-CAM++ trained on diverse data	100% sensitivity and specificity; significant specificity superiority over 11 experts	Provides an automated expert-committee second reader and may reduce unnecessary chest tube interventions
Objective 2 (BLUE-LUSS)	Cumbersome LUS protocols for ventilated infectious patients	Integration of BLUE points with cLUSS and qLUSS quantification	Excellent reliability (ICC > 0.92 for cLUSS); significant correlation with PaO2/FiO2 ratio	Enables rapid, single-operator, non-invasive bedside oxygenation monitoring
Objective 3 (automation and deployment)	Lack of real-time clinical AI integration and semantic mapping	U-Net segmentation with HAILO-accelerated Raspberry Pi 5 bedside runtime	87.53% segmentation accuracy; functional real-time bedside prototype	Bridges the gap between retrospective algorithms and deployable point-of-care software

Objective	Unmet need	Innovation	Main quantitative result	Translational relevance
Objective 4 (cadaveric model)	Absence of high-fidelity, safe PTX training models	Minimally invasive safety-triangle induction strategy	Educational utility kappa = 0.763; 826 PTX-positive AI M-mode images generated	Addresses data scarcity for AI training and provides realistic procedural simulation
Objective 5 (integration framework)	Lack of a standardized pathway from algorithm development to bedside use	Unified bedside framework combining acquisition standardization, Pi-local hardware, touchscreen UI, automation, and provenance-aware review	Functional manual and automation-enabled bedside runtime; integrated YOLO localisation, M-mode ensemble prediction, and structured export	Converts isolated AI components into a clinician-facing, auditable, deployable bedside decision-support system

Code and software availability

Software availability

The bedside Raspberry Pi runtime developed for the present thesis has been archived as a versioned public software snapshot on Zenodo (DOI: 10.5281/zenodo.19541650). The archived snapshot documents the translational bedside deployment branch of the project and includes the software components required to reproduce its architecture and workflow, including the Epiphan AV.io capture pathway, the Pi-local dashboard logic, the HAILO preparation utilities, and the automation-enabled bedside runtime. The public archive was intentionally limited to the software elements that are necessary for the thesis, and therefore does not include protected clinical datasets, unpublished internal development materials, or other non-essential assets outside the scope of the bedside runtime snapshot.

Transparent AI-assisted software development

During the software development and repository preparation phases of this work, the OpenAI Codex 5 family was used transparently and in accordance with the applicable rules for ethical AI-assisted academic work. This assistance was used for selected coding, refactoring, documentation, and repository-packaging tasks. All AI-assisted outputs were critically reviewed, tested, and curated by the author before integration into the final software snapshot or the present thesis. The scientific direction, implementation decisions, bedside validation, and full responsibility for the final software and manuscript remained with the author.