From Trauma CT to Pneumonia Risk

Deep learning-based quantification of pneumothorax, haemothorax and pulmonary contusion following blunt thoracic trauma

Vivian van Asperen

September 2026

Erasmus MC

From Trauma CT to Pneumonia Risk

Deep learning-based quantification of pneumothorax, haemothorax and pulmonary contusion following blunt thoracic trauma

V. F. van Asperen

Student number: 4839994

04-09-2026

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in *Technical Medicine*

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)

Dept. of Biomechanical Engineering, TU Delft
Dept. of Trauma Surgery, Erasmus MC
Biomedical Imaging Group Rotterdam, Erasmus MC

March 2026 – September 2026

Supervisor(s):

Dr. M.M.E. (Mathieu) Wijffels, MD, Erasmus MC
Dr. ir. T. (Theo) van Walsum, Erasmus MC

Thesis committee members:

Dr. M.M.E. (Mathieu) Wijffels, MD, Erasmus MC
Dr. ir. T. (Theo) van Walsum, Erasmus MC
Dr. S.A.G. (Sven) Meylaerts, MD, LUMC

Leids Universitair Medisch Centrum

TU Delft

Erasmus MC

Acknowledgements

Over the past months, I have greatly enjoyed my time at the Trauma Surgery Department of Erasmus MC and am very grateful to everyone who contributed to both this thesis and my experience throughout the project.

Mathieu, thank you for the trust and freedom you gave me throughout this project. I especially appreciated being able to explore the directions that interested me and decide where I wanted to invest my time, while always having the opportunity to discuss ideas with you when needed. Particularly during the early stages of the project, our discussions helped me shape the direction of this thesis. I am also grateful that you involved me in activities beyond my thesis, including meetings with external collaborators and working on an additional study together with Jonne. These opportunities brought welcome variation to my work and allowed me to experience different aspects of clinical research.

Theo, thank you for your guidance, particularly on the technical and methodological aspects of this project. I greatly appreciated the opportunities you gave me to present my work during the IGIT meetings. Being able to discuss my work in depth with other researchers challenged me to explain and defend my choices and taught me a great deal. I also really appreciated the enthusiasm with which you explained technical and methodological concepts. Your way of supervising often encouraged me to think through problems more carefully and arrive at solutions myself, which has been especially valuable throughout this thesis.

I would also like to thank all the students and PhD candidates at the department for making my time at Erasmus MC so enjoyable. The daily lunches and coffee breaks, as well as the occasional drinks at Westkop, certainly kept me motivated throughout the project.

Outside of Erasmus MC, I am grateful to my family, friends and Hiske for their support throughout these months. I greatly appreciated your interest in what I was working on, as well as the welcome distraction outside of it. Having so many enjoyable things to do and look forward to alongside my thesis made these past months all the more enjoyable.

Finally, I would like to thank all the surgeons and residents who participated in the pulmonary contusion observer study. I am well aware that repeatedly assessing pulmonary contusion on CT was not necessarily the most exciting addition to already busy working days. I therefore greatly appreciate the time and effort you were willing to put into the assessments, which formed an essential part of this thesis.

V. F. van Asperen
Rotterdam, September 2026

AI statement

During the preparation of this thesis, AI-based tools were used to support language editing, improve the clarity and readability of the text, assist with programming-related questions, and support the formatting of the LaTeX layout. The tools used included ChatGPT, Claude, and Quillbot. All AI-generated suggestions were critically reviewed, verified, and adapted before being incorporated into the thesis. Data collection, statistical analyses, methodological decision-making, interpretation of results, and formulation of conclusions were conducted by the author.

Abstract

Introduction: Blunt thoracic trauma frequently results in pneumothorax (PTX), haemothorax (HTX), and pulmonary contusion (PC), injuries that may contribute to post-traumatic pulmonary complications such as pneumonia. Although computed tomography (CT) enables detailed assessment of these injuries, their extent is generally not quantified routinely. Automated image analysis may enable objective quantification of thoracic injury burden and provide quantitative imaging features for prediction of post-traumatic pneumonia. This study investigated whether PTX, HTX, and PC can be automatically and reliably quantified from admission CT and whether these quantitative imaging features can be integrated into pneumonia prediction.

Methods: PTX and HTX were segmented using a multi-class nnU-Net model with lung parenchyma as an auxiliary anatomical class. Model configurations and post-processing thresholds were optimised during development, and the final pipeline was evaluated on an independent test set. For PC, six clinical observers assessed contusion extent within twelve predefined lung regions. Inter- and intraobserver agreement were evaluated, and the mean assessment of the six observers was used as a consensus reference for development of a 3D convolutional neural network (CNN) for automated PC quantification. Finally, the automatically derived PTX, HTX, and PC measurements were evaluated as radiological features within an existing clinical–radiological model for post-traumatic pneumonia prediction.

Results: On independent testing, PTX achieved a sensitivity of 1.00, specificity of 0.80, and mean absolute volume error of 5.2 mL. For HTX, sensitivity was 0.81, specificity 0.82, and mean absolute volume error 35.4 mL. Including lung parenchyma as an auxiliary class significantly reduced HTX volume error and false-positive predictions. PC assessment showed moderate interobserver agreement and substantial variation in intraobserver agreement. Automated PC quantification achieved a patient-level mean absolute error of 5.70 percentage points and a calibration slope of 0.68 on independent testing, with better performance at patient than regional level. Integration of the resulting automated pulmonary injury measurements into the pneumonia prediction framework yielded a test area under the receiver operating characteristic curve (AUC) of 0.81, compared with 0.83 for the previously developed combined clinical–radiological model.

Conclusion: Automated quantification of pulmonary injury from trauma CT was feasible, although performance differed between injury types. The resulting PTX, HTX, and PC measurements could be integrated into the existing pneumonia prediction framework while largely preserving discrimination compared with the original combined model. External validation is required to determine whether these measurements and resulting risk estimates can provide clinically meaningful support for patient management.

Contents

1

Introduction

1.1. Clinical background

Blunt thoracic trauma is a common mechanism of injury in trauma care, accounting for approximately 10–25% of all trauma cases [1, 2, 3, 4, 5]. Thoracic injuries such as rib fractures, pneumothorax, haemothorax and pulmonary contusion are frequent findings following blunt thoracic trauma and may substantially affect respiratory function [6, 7]. Collectively, these injuries are associated with an increased risk of pulmonary complications, including pneumonia [6, 8, 9, 10, 11]. Pneumonia is an important determinant of adverse clinical outcome and has been identified as an independent predictor of in-hospital mortality in trauma patients [12]. In a meta-analysis, pneumonia was the complication most strongly associated with mortality following blunt chest wall trauma [13].

Post-traumatic pneumonia typically develops several days after the initial traumatic insult, which complicates risk assessment at presentation and limits the opportunity for timely preventive intervention [14]. Early identification of patients at increased risk may therefore support closer respiratory monitoring and preventive management. In patients with rib fractures, early risk stratification may also influence treatment planning. Surgical stabilisation can improve selected clinical outcomes, although optimal patient selection remains uncertain [15, 16].

The development of pneumonia after blunt thoracic trauma is influenced by both patient-related and injury-related factors. Thoracic injuries, including rib fractures, pneumothorax and haemothorax, may impair respiratory function through pain, reduced chest wall movement, compromised lung expansion and impaired regional ventilation. These changes can contribute to atelectasis and impaired secretion clearance [8, 9, 10, 11]. Pulmonary contusion represents direct injury to the lung parenchyma and has been associated with ventilator-associated pneumonia and adverse pulmonary outcomes [6, 8]. Quantifying pneumothorax and haemothorax volume and pulmonary contusion extent may therefore complement established clinical predictors by providing objective measures of thoracic injury burden.

The potential clinical value of these measurements is supported by associations between injury burden and outcome. Pulmonary contusion involving approximately 18–24% or more of total lung volume has been associated with worse in-hospital outcomes [17], while CT-based haemothorax volume has been associated with the need for drainage intervention [18]. These findings suggest that quantitative imaging measures may provide clinically relevant information beyond conventional categorical variables. However, evidence on the prognostic value of pneumothorax volume for post-traumatic pneumonia remains limited.

In routine clinical practice, pneumothorax, haemothorax and pulmonary contusion are generally assessed visually rather than through explicit volumetric delineation. Obtaining quantitative measurements by manual or semi-manual CT segmentation is time-consuming and requires anatomical expertise [19]. Reproducible quantification may also be challenging: small pneumothoraces and haemothoraces can be difficult to delineate [20, 21], while pulmonary contusions often have ill-defined borders and may overlap with atelectasis, aspiration, dependent density or other parenchymal abnormalities on CT [6, 17]. These challenges may contribute to inter- and intraobserver variability, which is particularly relevant for pulmonary contusion assessment and may affect the construction of reliable reference standards for supervised deep learning [22]. Automated CT-based segmentation may therefore offer a more objective and reproducible approach for extracting quantitative thoracic imaging features [23].

This study therefore aimed to investigate whether thoracic injury burden can be automatically and reliably quantified from admission CT and whether these quantitative imaging features provide additional information for prediction of post-traumatic pneumonia following blunt thoracic trauma.

1.2. Previous work on automated pulmonary injury quantification

Automated CT-based methods have been developed for quantification of pneumothorax, haemothorax and pulmonary contusion, but evaluation in representative trauma populations remains limited. A systematic review conducted as part of the present research identified substantially more studies addressing pneumothorax than haemothorax or pulmonary contusion. Deep learning-based pneumothorax segmentation has shown promising performance, with reported Dice similarity coefficients of approximately 0.87–0.96 [24, 25, 23, 26, 27, 28]. However, several studies included exclusively pneumothorax-positive patients, while negative cases and concurrent thoracic abnormalities were handled inconsistently [24, 25, 23, 28].

For haemothorax, Dreizin et al. reported a mean Dice coefficient of 0.75 and strong volumetric agreement using an nnU-Net-based approach, but trace haemothoraces of 50 mL or less were excluded [21]. Quantification of pulmonary contusion has been investigated using deep learning-based segmentation and attenuation-based or hybrid approaches [29, 30, 31], but substantial heterogeneity exists in how contusion extent is defined and quantified.

An additional challenge in pulmonary contusion quantification concerns the reference annotations used for model development. Previous studies used different annotation strategies and levels of expert review, while formal assessment of inter- or intraobserver variability was not reported [29, 30, 31]. Further standardisation of pulmonary contusion quantification and evaluation of inter- and intraobserver agreement have previously been highlighted as areas for future research [17]. More broadly, existing automated pulmonary injury methods have predominantly been evaluated in retrospective single-centre datasets without external validation.

1.3. Previous work on post-traumatic pneumonia prediction

Several prediction models have been developed to estimate the risk of pneumonia or pulmonary complications after blunt thoracic trauma [32, 33, 34]. These models primarily combine clinical and injury-related variables, including age, injury severity, and rib fracture characteristics. Some studies have additionally incorporated CT-based measures of pulmonary contusion severity. These include the BPC18 score, which grades contusion density across six lung regions, and Yang's index, which approximates pulmonary contusion volume from three-dimensional CT measurements [33, 35]. However, quantitative measures of pleural injury extent, and automated measurements of pulmonary contusion extent, remain limited in existing prediction models.

Previous work within the present research line established an automated CT-based framework for thoracic injury characterisation. Marting et al. developed a deep learning-based method for automatic detection and classification of rib fractures on chest CT [36]. These automatically derived rib fracture features were subsequently combined with clinical variables and other automatically derived CT characteristics in a machine-learning model for post-traumatic pneumonia prediction developed using the same research cohort [37]. However, pneumothorax, haemothorax and pulmonary contusion were quantified using previously available image-analysis methods rather than dedicated deep learning models developed and independently evaluated for quantitative pulmonary injury assessment.

1.4. Contributions

Building on the methodological and clinical gaps outlined above, the present study makes four main contributions:

1. A deep learning-based multi-class segmentation pipeline for quantitative assessment of pneumothorax and haemothorax, including systematic evaluation of model configuration, post-processing thresholds and the contribution of lung parenchyma segmentation.
2. A multi-observer consensus reference for regional pulmonary contusion assessment, supported by quantitative evaluation of inter- and intraobserver agreement.
3. A three-dimensional convolutional neural network for automated regional pulmonary contusion quantification, together with evaluation of the influence of training strategy and dataset composition.
4. Integration of automatically derived pneumothorax, haemothorax and pulmonary contusion measurements into an existing pneumonia prediction framework, enabling evaluation of their performance on a held-out test set.

The remainder of this study is organised as follows. Chapter 2 describes the study population, datasets and the methods developed. Chapter 3 presents the experiments and results for automated pneumothorax and haemothorax quantification, pulmonary contusion consensu reference construction and automated quantification, and integration of the resulting imaging features into the pneumonia prediction model. Chapter 4 discusses the principal findings, methodological considerations, clinical implications and limitations, followed by the overall conclusion in Chapter 5.

2

Methods

2.1. Study framework and workflow

An overview of the study framework and integration of the different imaging features is illustrated in Figure 1. Pneumothorax, haemothorax and pulmonary contusion were quantified from admission chest CT scans using dedicated automated image analysis models developed in this study, while rib fracture characteristics were obtained using an existing automated pipeline. For pulmonary contusion, observer agreement was evaluated before construction of the consensus reference and model development. The resulting imaging features were subsequently combined with clinical variables and evaluated within an existing machine-learning framework for prediction of post-traumatic pneumonia.

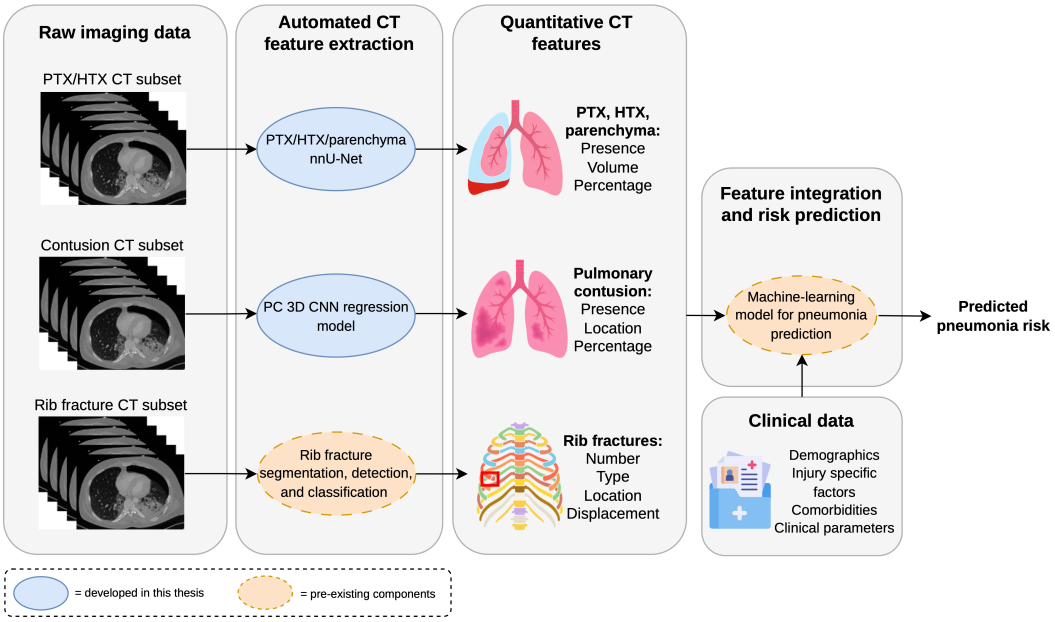


Figure 1: Overview of the study framework and automated imaging-to-prediction workflow. CNN, convolutional neural network; CT, computed tomography; HTX, haemothorax; PC, pulmonary contusion; PTX, pneumothorax.

2.2. Data

2.2.1. Study population

The dataset underlying the pneumonia prediction tool was derived from the previously conducted RIB-Spiro and RIB-CPR cohort studies at Erasmus MC, whose original eligibility criteria and study procedures have been described previously [38, 39]. For the present study, patients from these cohorts were eligible for inclusion if they were aged 16 years or older, had at least one CT-confirmed rib fracture, had complete electronic health records available, and had at least 30 days of follow-up for pneumonia assessment. Exclusion criteria were a pre-existing pulmonary infection at the time of trauma, an unavailable or missing CT scan, and incomplete visualization of the ribs.

2.2.2. Study design and dataset partitioning

The source radiological dataset had previously been partitioned for development of the final pneumonia prediction model. The predefined pneumonia test set was kept completely separate and was not used for development or evaluation of the automated imaging models.

The remaining non-test data were used for development of the pulmonary injury models. Five-fold cross-validation within the predefined training set was used for model development, internal validation and model selection. The predefined validation set served as a held-out test set for evaluation of the pulmonary injury models before integration into the pneumonia prediction framework (Figure 2).

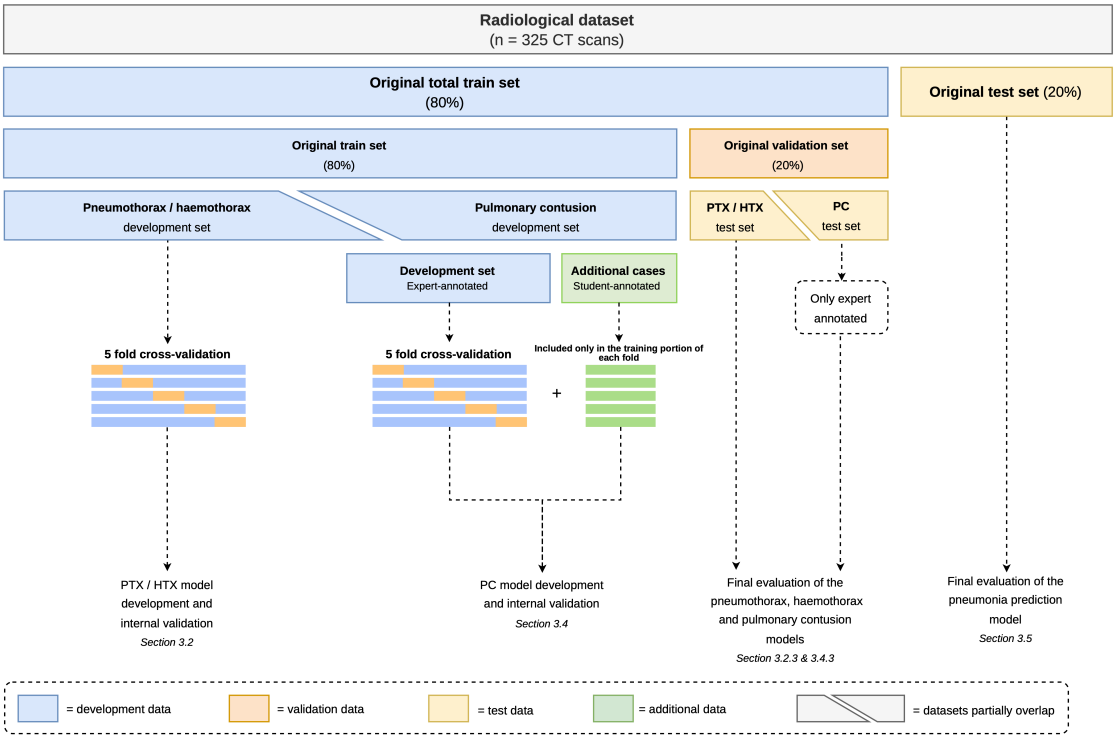


Figure 2: Overview of the study design and dataset partitioning strategy. CT, computed tomography; HTX, haemothorax; PC, pulmonary contusion; PTX, pneumothorax.

2.2.3. Dataset selection

For each automated imaging task, dedicated subsets were selected from the available non-test data. Positive cases were selected to include a broad range of injury severities, while negative cases were included to support discrimination between injured and non-injured scans. Scans with concurrent thoracic abnormalities, including atelectasis, bullae, emphysema and other traumatic lung abnormalities, were intentionally retained because these findings could resemble the target abnormalities and therefore represent clinically relevant sources of error. The pneumothorax/haemothorax and pulmonary contusion datasets were selected independently for their respective modelling tasks and therefore partially overlapped.

2.2.4. Pneumonia outcome

The primary outcome was pneumonia diagnosed within 30 days after trauma. Pneumonia was identified from the clinical records based on clinical signs, including fever, respiratory distress and increased sputum production; laboratory findings including leukocytosis and elevated C-reactive protein; and radiological evidence of a new pulmonary infiltrate, consistent with the Centers for Disease Control and Prevention criteria [40].

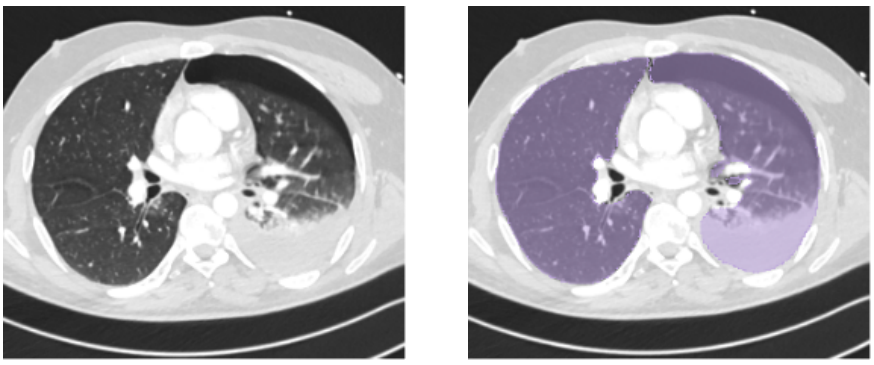
2.3. Reference annotations

2.3.1. Pleural injury reference annotations

Voxel-level reference segmentations were created for pneumothorax, haemothorax and lung parenchyma. Pneumothorax was defined as air located within the pleural space, whereas haemothorax comprised blood or blood-containing fluid within the pleural space. Lung parenchyma was defined as all pulmonary tissue within the visceral pleural boundary, irrespective of its degree of aeration, thereby including contused, consolidated and atelectatic lung regions.

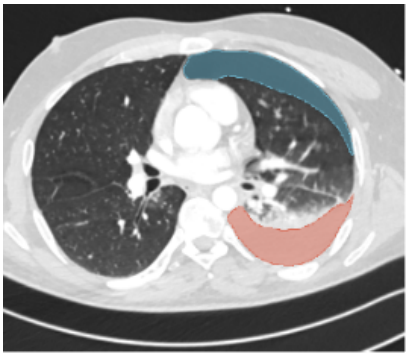
Reference annotations were created in 3D Slicer by a trained Technical Medicine master's student. Ambiguous cases were reviewed together with a trauma surgeon with more than ten years of clinical experience in the treatment of thoracic injuries. Pneumothorax and haemothorax were manually delineated on selected axial slices, followed by interpolation and manual correction where necessary. Particular attention was paid to distinguishing pneumothorax from emphysematous bullae and haemothorax from atelectatic, contused or densely consolidated lung tissue. Preliminary single-label models were additionally used as a quality-control step to identify potential annotation inconsistencies, which were reviewed and corrected where necessary.

Previously generated thoracic cavity masks obtained with TotalSegmentator were available for all scans [41]. These masks comprised the volume enclosed by the parietal pleural boundary, including lung parenchyma and any pleural air or fluid, while excluding the mediastinum and chest wall. Lung parenchyma masks were obtained by subtracting the manually annotated pneumothorax and haemothorax segmentations from the thoracic cavity mask. The anatomical reference volumes are illustrated in Figure 3.

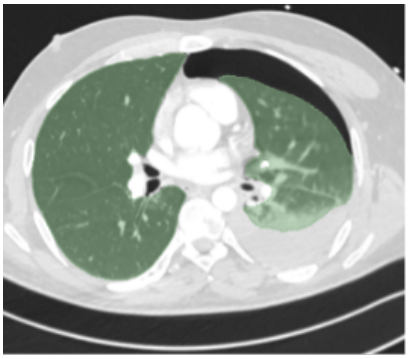


(a) Original CT

(b) Thoracic cavity



(c) Pleural injuries



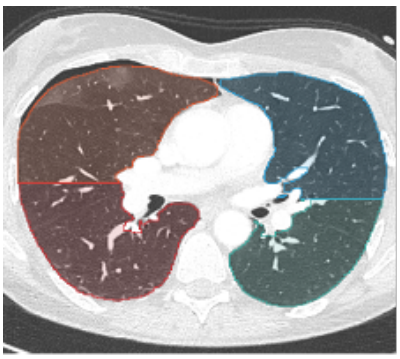
(d) Lung parenchyma

Figure 3: Determination of anatomical reference volumes for quantitative injury assessment. (a) Original CT image. (b) Thoracic cavity mask obtained using TotalSegmentator. (c) Manually annotated pleural injuries, comprising pneumothorax (blue) and haemothorax (red). (d) Derived lung parenchyma mask (green), obtained by subtracting PTX and HTX from the thoracic cavity mask. CT, computed tomography; HTX, haemothorax; PTX, pneumothorax.

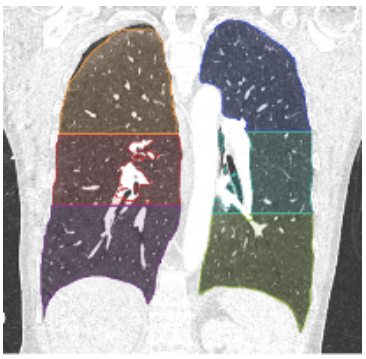
Thoracic cavity masks served as anatomical reference volumes for expressing pneumothorax and haemothorax volumes relative to thoracic cavity volume, whereas the derived lung parenchyma masks served as reference volumes for expressing pulmonary contusion extent. For pulmonary contusion, 100% was defined as the segmented lung parenchyma present on the admission CT.

2.3.2. Pulmonary contusion reference annotations

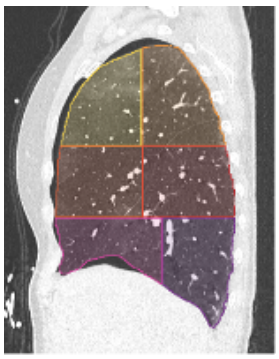
Pulmonary contusion severity was quantified as the estimated percentage of affected lung parenchyma within twelve predefined anatomical lung regions. Each lung was automatically subdivided into six equal-volume regions, allowing patient-level pulmonary contusion severity to be calculated as the unweighted mean across all twelve regions. Colour-coded overlays were generated in 3D Slicer to facilitate standardised assessment (Figure 4).



(a) Axial



(b) Coronal



(c) Sagittal

Figure 4: Automatic subdivision of the lung parenchyma into twelve anatomical regions used for standardised pulmonary contusion assessment. The same anatomical partitioning is shown in the axial (a), coronal (b) and sagittal (c) planes.

Forty-seven CT scans were independently assessed by six clinical observers, comprising four trauma surgeons and two surgical residents with varying levels of experience in the management of patients with thoracic trauma. Observers estimated the pulmonary contusion percentage within each anatomical region using a continuous scale ranging from 0% to 100% and were encouraged to use rounded estimates rather than overly precise values. The complete assessment protocol is provided in Appendix B.

To evaluate intraobserver agreement, four observers repeated the assessment after a minimum interval of two weeks using the same assessment protocol.

For the pulmonary contusion model, the reference value for each region was defined as the mean pulmonary contusion percentage across the six clinical observers during the first assessment round. Inter- and intraobserver agreement were evaluated in Sections 3.3.1 and 3.3.2.

2.4. Model development

2.4.1. Pleural injury model

Pneumothorax and haemothorax segmentation was implemented using nnU-Net v2 with PyTorch 2.11.1 in Python 3.11.3. nnU-Net automatically configures preprocessing, image resampling, network architecture and training parameters based on the characteristics of the input dataset [42]. Unless stated otherwise, the automatically generated configuration and default training settings were used. Model development was performed using five-fold cross-validation on the development set.

The evaluated configurations included two-dimensional and three-dimensional full-resolution models, as well as an ensemble of both configurations. During inference, CT scans were cropped to the thoracic field of view and the resulting predictions were restored to the original image space before volumetric feature extraction.

2.4.2. Pulmonary contusion model

A separate model input was generated for each of the twelve predefined anatomical lung regions described in Section 2.3.2.

For each region, the smallest three-dimensional bounding box containing the regional mask was extracted. A fixed input shape was determined from the maximum regional bounding-box dimensions across the dataset, with an additional ten-voxel margin on each side. Smaller CT blocks and masks were zero-padded to $254 \times 241 \times 214$ voxels. No spatial resampling or mask erosion was applied.

CT intensities were clipped to -1024 to 400 Hounsfield units (HU) and scaled to 0–1. Voxels outside the selected anatomical region were set to zero after normalisation. The CT block and binary regional mask formed a two-channel input, while anatomical region identity was encoded as a twelve-dimensional one-hot vector.

The compact three-dimensional convolutional neural network (CNN) comprised four convolutional blocks with 8, 16, 32 and 64 output channels. Each block consisted of a three-dimensional convolution, batch normalisation and rectified linear unit activation, with max pooling after the first three blocks. Global adaptive average pooling generated a 64-dimensional image feature vector, which was concatenated with the region identifier. The resulting 76-dimensional vector was passed through a 32-unit fully connected layer with ReLU activation and dropout of 0.2, followed by a linear output layer. Predictions were clipped to 0–100% before evaluation. A compact architecture was selected to limit model complexity and reduce the risk of overfitting given the limited number of annotated patients. The resulting model input and network architecture are illustrated in Figure 5.

Models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. A ReduceLROnPlateau scheduler was used with a reduction factor of 0.5 and a minimum learning rate of 1×10^{-6}. Early stopping was based on the validation value of the respective training objective, with a patience of 15 epochs.

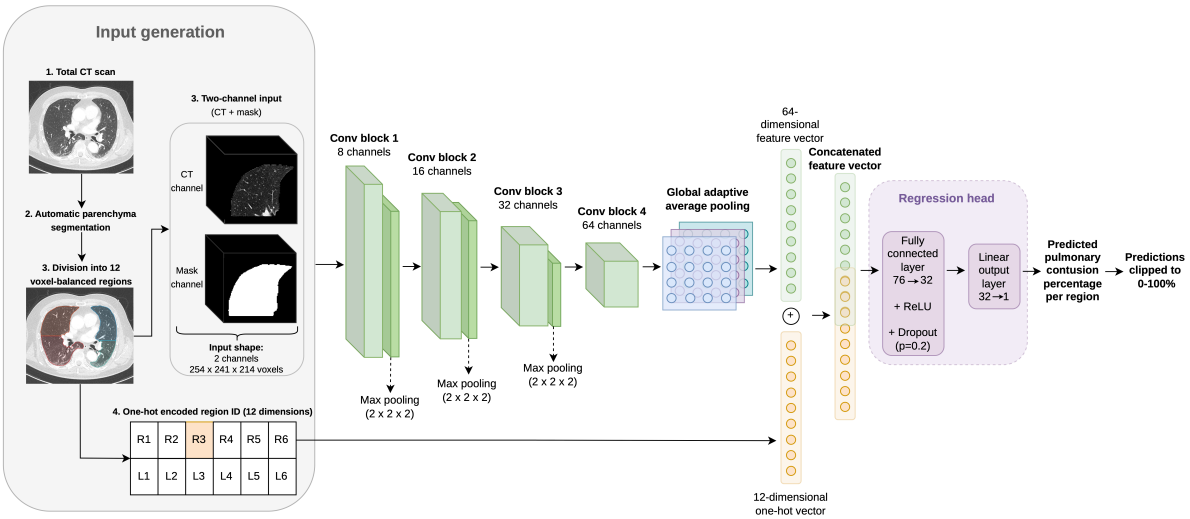


Figure 5: Input generation and network architecture of the pulmonary contusion convolutional neural network (CNN).

2.4.3. Pneumonia prediction

Pneumonia prediction was performed using an existing machine-learning pipeline. The original training, validation and test partitioning was retained. Candidate predictors consisted of clinical variables and automatically derived CT features. Categorical variables were encoded and continuous variables were normalised prior to model development.

Feature selection was performed using Least Absolute Shrinkage and Selection Operator (LASSO) regression. Alternative representations of the same injury characteristic were defined as mutually exclusive feature groups to reduce redundancy. Within each group, only the feature with the largest absolute LASSO coefficient was retained.

Five classification algorithms were evaluated: Naive Bayes, logistic regression, random forest, support vector machine and extreme gradient boosting (XGBoost). Classifier hyperparameters were optimised using stratified five-fold cross-validation on the training set, with balanced accuracy as the optimisation criterion. The validation set was reserved for final model selection based on area under the receiver operating characteristic curve (AUC), while the independent test set remained excluded from model development and selection. Following final model selection, the selected configuration was retrained on the combined training and validation sets and evaluated on the independent test set.

2.5. Evaluation

2.5.1. Pleural injury model evaluation strategy

Because the primary objective of the pleural injury models was accurate quantification of pneumothorax and haemothorax volume for downstream pneumonia prediction, volumetric accuracy was considered the primary evaluation outcome. Segmentation accuracy was evaluated as a complementary measure of spatial overlap.

All primary performance analyses were conducted at the scan level. Model performance was evaluated separately for pleural injuries and lung parenchyma. For pneumothorax and haemothorax, detection performance was assessed per scan using sensitivity, specificity and balanced accuracy after application of the selected post-processing thresholds. Quantitative performance was evaluated by comparing predicted and reference volumes, expressed in millilitres and as percentages of thoracic cavity volume. Mean absolute error (MAE) was calculated across all scans to reflect overall model performance and separately across reference-positive scans to assess volumetric accuracy when injury was present. Bias was defined as the mean signed volume error. Spatial overlap was assessed using the Dice similarity coefficient in reference-positive scans only. Lung parenchyma segmentation was evaluated using the Dice similarity coefficient and absolute volume error.

For the ablation analysis, paired differences in absolute volume error between the models with and without the lung parenchyma class were evaluated using the Wilcoxon signed-rank test.

Finally, scan-level error analysis was performed to characterise false-positive and false-negative predictions and to identify systematic patterns in volumetric over- or underestimation.

2.5.2. Inter- and intraobserver agreement evaluation strategy

Analyses were performed at both the segment and patient levels. Patient-level pulmonary contusion severity was obtained by averaging the pulmonary contusion percentages across the twelve anatomical lung regions for each observer. Interobserver agreement was assessed across the six clinical observers, whereas intraobserver agreement was evaluated for the four observers who repeated all assessments after a minimum interval of two weeks.

Interobserver agreement was quantified using a two-way absolute agreement intraclass correlation coefficient for single measures, ICC(A,1), based on the first-round ratings of all six clinical observers. Intraobserver agreement was evaluated separately for each of the four observers who repeated the assessment, using ICC(A,1) to quantify absolute agreement between that observer's first- and second-round ratings. According to the classification proposed by Koo and Li, ICC values below 0.50 were interpreted as poor, values between 0.50 and 0.75 as moderate, values between 0.75 and 0.90 as good, and values above 0.90 as excellent [43].

Because ICC does not directly describe the magnitude or direction of disagreement, intraobserver analyses were complemented by the mean absolute error, mean signed difference between repeated assessments, 95% Bland–Altman limits of agreement and the percentage of observations differing by no more than 10 percentage points. As an exploratory analysis, interobserver and intraobserver disagreement were additionally evaluated across pulmonary contusion severity groups. Detailed definitions and calculation procedures are provided in Appendix C.

2.5.3. Pulmonary contusion model evaluation strategy

Patient-level MAE was considered the primary performance metric because it reflects the overall error in pulmonary contusion estimation for each patient. Segment-level MAE was evaluated as a secondary measure to assess prediction accuracy within individual anatomical lung regions. Additional performance measures included the Pearson correlation coefficient, calibration slope, prediction range and prediction standard deviation to evaluate linear association with the reference, calibration and the ability of the model to reproduce the full spectrum of pulmonary contusion severities.

2.5.4. Pneumonia prediction model evaluation strategy

Because the primary objective of the pneumonia prediction model was accurate discrimination between patients with and without post-traumatic pneumonia, the area under the receiver operating characteristic curve (AUC) was considered the primary performance metric. Balanced accuracy, sensitivity, specificity, positive predictive value and negative predictive value were additionally calculated to characterise classification performance.

Potential overfitting was assessed by comparing training

and cross-validation balanced accuracy. Feature contributions to the final model predictions were evaluated using SHapley Additive exPlanations (SHAP).

For all statistical analyses, statistical significance was defined as $p < 0.05$.

3

Experiments and Results

3.1. Datasets

3.1.1. Pleural injury dataset

The pleural injury dataset comprised 157 development and 33 independent test CT scans, selected according to the partitioning strategy described in Section 2.2.2. The development set included scans with pneumothorax, haemothorax, both injuries and neither injury. The independent test set was used exclusively for final evaluation. Dataset composition is presented in Table 1.

Table 1: Distribution of pneumothorax and haemothorax in the pleural injury datasets.

Characteristic	Development	Independent test
Total	157	33
PTX only	30 (19%)	4 (12%)
HTX only	27 (17%)	7 (21%)
Both injuries	40 (25%)	9 (27%)
Neither injury	60 (38%)	13 (39%)

HTX, haemothorax; PTX, pneumothorax.

3.1.2. Pulmonary contusion dataset

The pulmonary contusion dataset comprised 70 CT scans. Of 47 scans assessed by the six clinical observers, 37 were included in the development set and ten were reserved for independent testing. An additional 23 scans were assessed by the student following the validation described in Section 3.4.1 and were used for training only. Dataset composition is presented in Table 2.

Five-fold cross-validation was performed using the expert-annotated development cases. Student-annotated cases were included only in the training portion of each fold and were excluded from validation, model selection or independent testing.

Table 2: Pulmonary contusion annotation distribution.

Dataset	Expert	Student	Total
Development set	37	23	60
Independent test set	10	0	10

3.1.3. Pneumonia dataset

The original pneumonia prediction dataset comprised 325 patients with complete data. One patient was excluded because the corresponding CT scan could not be processed, resulting in 324 patients. The original training, validation and independent test partitioning was retained. Dataset composition is presented in Table 3.

Table 3: Pneumonia outcome distribution across the dataset partitions.

Dataset	No pneumonia	Pneumonia	Total
Training set	184	24	208
Validation set	46	7	53
Test set	57	6	63
Total	287	37	324

3.2. Pleural injury segmentation

3.2.1. Ablation of the parenchyma label

Pneumothorax and haemothorax were combined into a single multi-class segmentation model because both injuries occupy the pleural space and frequently coexist. Lung parenchyma was included as an auxiliary class to provide explicit anatomical information on the boundary between pulmonary tissue and the pleural cavity.

The contribution of the lung parenchyma class was evaluated in a matched ablation experiment. The final pipeline was retrained without this additional class, while keeping the dataset, data splits, network configuration, training settings and evaluation strategy unchanged.

Table 4: Contribution of the lung parenchyma class to pleural injury segmentation performance.

	PTX		HTX	
Metric	**Without**	**With**	**Without**	**With**
MAE, all cases (mL)	6.0	5.2	56.6	35.4
Sensitivity	1.00	1.00	0.81	0.81
Specificity	0.75	0.80	0.59	0.82
Difference in absolute volume error (95% CI)				*p*
PTX	-0.75 mL (-2.50 to 0.84)			0.277
HTX	-21.19 mL (-43.11 to -4.80)			0.009

CI, confidence interval; HTX, haemothorax; MAE, mean absolute error; PTX, pneumothorax.

The addition of the lung parenchyma class had limited effect on pneumothorax performance, whereas haemothorax volume error was significantly reduced and specificity improved (Table 4). The three-label representation was therefore retained for the final pipeline.

3.2.2. Segmentation configuration and threshold optimisation

Three nnU-Net configurations and ten thresholds per injury were evaluated, resulting in 300 candidate pipelines. Thresholding was applied to suppress small false-positive predictions and to prevent very small predicted volumes from being interpreted as clinically relevant injuries. The evaluated ranges were centred around volume thresholds previously used in pneumothorax and haemothorax segmentation studies, approximately 1 mL and 50 mL, respectively [24, 25]. The evaluated threshold values and additional details of the model-selection analysis are provided in Appendix A.

Because the intended output was quantitative injury volume for downstream pneumonia prediction, pipeline selection prioritised volumetric accuracy rather than voxel-wise overlap alone. Candidate pipelines were first required to achieve predefined sensitivities of at least 0.95 for pneumothorax and 0.90 for haemothorax. The pneumothorax threshold was selected to approximate the high sensitivity reported for automated CT-based pneumothorax detection [44]. A slightly lower threshold was prespecified for haemothorax because it represents a more challenging detection and segmentation task, while still prioritising injury detection before optimisation of volumetric accuracy.

Eligible pipelines were ranked according to the mean of the pneumothorax and haemothorax MAEs. In the event of an exact tie, the pipeline with the highest mean balanced accuracy was selected, followed by the pipeline with the lower post-processing thresholds. The selection procedure is summarised in Figure 6.

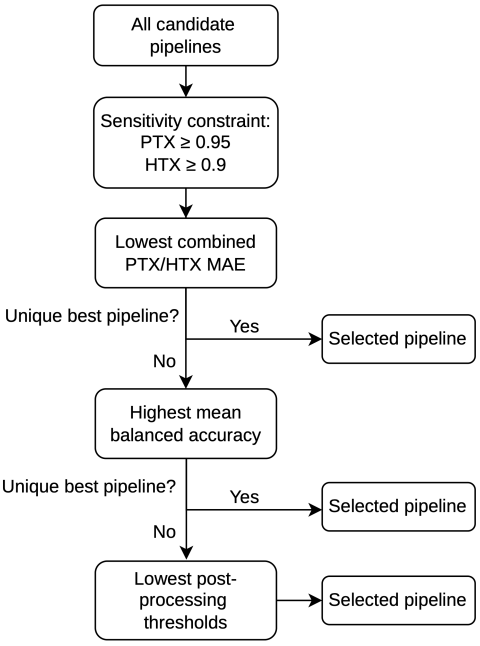


Figure 6: Hierarchical model selection strategy for the pleural injury segmentation pipeline. HTX, haemothorax; MAE, mean absolute error; PTX, pneumothorax.

Of the 300 available pipelines, 103 satisfied both sensitivity requirements. The three-dimensional full-resolution configuration with thresholds of 3 mL for pneumothorax and 40 mL for haemothorax achieved the lowest combined MAE and was selected for independent evaluation.

3.2.3. Independent test performance

Independent test-set performance is summarised in Table 5. Pneumothorax showed higher spatial overlap and smaller volume errors than haemothorax, while haemothorax showed an overall tendency towards volume underestimation. Lung parenchyma segmentation showed high spatial overlap, supporting its subsequent use as the anatomical reference volume for pulmonary contusion quantification.

Table 5: Independent test-set performance of the pleural injury segmentation pipeline.

Metric	PTX	HTX	Parenchyma
Sensitivity	1.00	0.81	–
Specificity	0.80	0.82	–
Dice, reference-positive	0.82	0.61	0.98
MAE, all cases (mL)	5.20	35.40	31.30
MAE, reference-positive (mL)	10.70	62.80	–
MAE, all cases (% thoracic cavity volume)	0.12	0.91	0.81
Bias, reference-positive (mL)	+6.70	−41.80	+2.20

Dice was calculated in reference-positive scans for PTX and HTX. HTX, haemothorax; MAE, mean absolute error; PTX, pneumothorax.

Reference and predicted injury volumes for reference-positive cases are shown in Figure 7. Predictions generally followed the line of identity, with greater deviations observed for haemothorax, particularly at larger reference volumes.

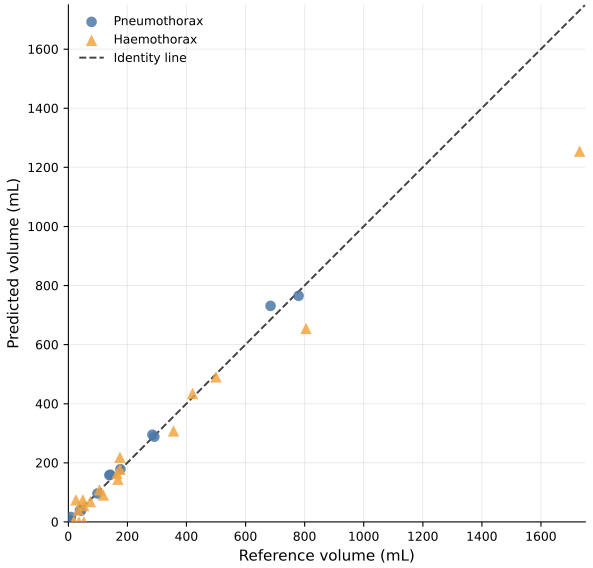


Figure 7: Reference and predicted pleural injury volumes on the independent test set for reference-positive pneumothorax and haemothorax cases. The dashed line represents perfect agreement.

3.2.4. Error analysis

False-negative detections occurred exclusively in small haemothoraces. In all three cases, the predicted haemothorax volume was below the 40-mL post-processing

threshold and was therefore classified as negative. Three false-positive haemothorax predictions were observed, with predicted volumes ranging from 43.0 to 75.1 mL. In contrast, the two largest haemothoraces were both correctly detected, although their volumes were underestimated. One case showed a large underestimation of 476.4 mL, which contributed substantially to the overall negative volume error (Figure 7).

Four false-positive pneumothorax predictions were observed, with predicted volumes ranging from 3.3 to 15.0 mL.

Visual review of the false-positive predictions revealed several recurring sources of misclassification. Pneumothorax false positives primarily involved intrapulmonary or subdiaphragmatic air that was incorrectly classified as pleural air, including emphysematous bullae, a focal low-attenuation parenchymal region, and colonic gas below the diaphragm in a scan affected by motion artefact. Haemothorax false positives predominantly involved increased-density pulmonary abnormalities that were incorrectly classified as pleural fluid, with the predicted regions extending through the lung parenchyma rather than conforming to the pleural space.

3.3. Pulmonary contusion assessment agreement

3.3.1. Interobserver agreement

The interobserver agreement analysis included 47 patients (564 lung segments) scored by six clinical observers. Agreement between individual observer ratings was moderate at both the segment and patient level (Table 6).

Table 6: Interobserver agreement for pulmonary contusion assessment.

Metric	Segment	Patient
ICC(A,1)	0.654	0.635
95% CI	0.55–0.74	0.47–0.76
Interpretation	Moderate	Moderate

CI, confidence interval; ICC, intraclass correlation coefficient.

Interobserver disagreement increased with pulmonary contusion severity. Cases without pulmonary contusion showed almost no disagreement, whereas the largest disagreement was observed in segments with pulmonary contusion scores of 30% or higher (Figure 8). Exact descriptive results by severity category are provided in Appendix C.

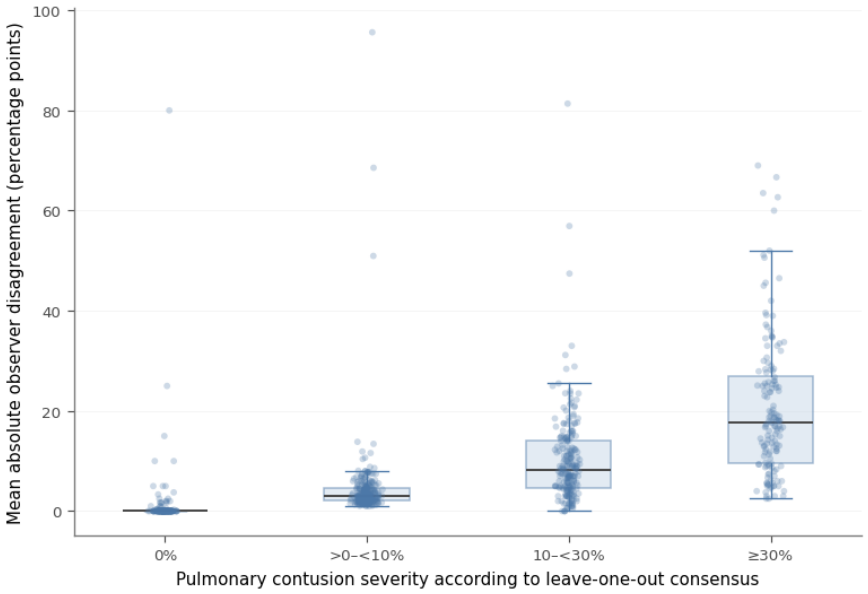


Figure 8: Interobserver disagreement across pulmonary contusion severity categories. Disagreement is shown as the absolute difference between individual observer scores and the leave-one-out consensus.

3.3.2. Intraobserver agreement

Observer A showed excellent intraobserver agreement at both the segment and patient levels. Agreement was lower for observers B, C and D, particularly for observer B at the patient level (Table 7). Mean absolute errors were smaller at the patient level for all four observers. Bland–Altman analysis showed small systematic biases for all observers, but substantially greater variability between repeated assessments for observers C and D than for observers A and B (Table 7; Figure 9).

Table 7: Intraobserver agreement statistics for repeated pulmonary contusion assessments.

Observer	Level	ICC(A,1) (95% CI)	Interpretation	MAE (pp)	Bias (pp)	LoA (pp)	Within ±10 pp (%)
A	Segment	0.922 (0.89–0.95)	Excellent	4.15	0.97	−14.4 to 16.3	91.67
A	Patient	0.961 (0.93–0.98)	Excellent	2.17	0.97	−5.3 to 7.3	97.87
B	Segment	0.551 (0.30–0.80)	Moderate	6.37	−1.93	−30.3 to 26.4	86.17
B	Patient	0.416 (0.14–0.87)	Poor	4.67	−1.93	−24.5 to 20.6	89.36
C	Segment	0.741 (0.60–0.85)	Moderate	6.51	−3.73	−30.8 to 23.3	87.41
C	Patient	0.692 (0.49–0.88)	Moderate	5.57	−3.73	−20.9 to 13.4	85.11
D	Segment	0.592 (0.39–0.79)	Moderate	13.15	−4.78	−53.7 to 44.1	75.18
D	Patient	0.521 (0.27–0.81)	Moderate	10.38	−4.78	−45.2 to 35.6	76.60

CI, confidence interval; ICC, intraclass correlation coefficient; LoA, limits of agreement; MAE, mean absolute error; pp, percentage points.

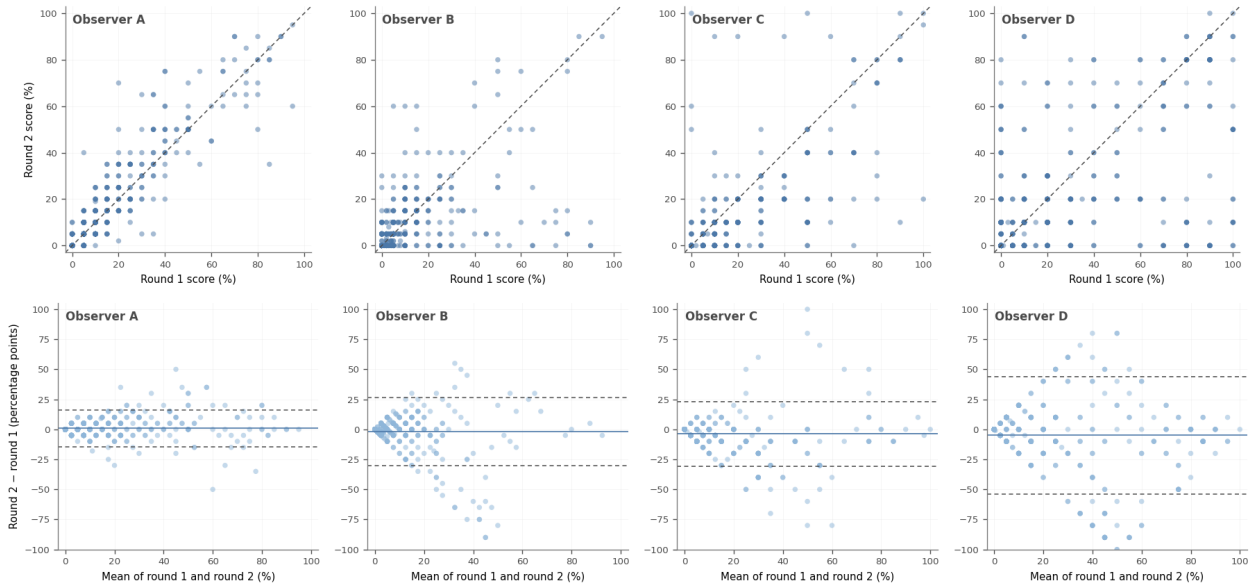


Figure 9: Scatterplots and Bland–Altman plots for intraobserver agreement between repeated pulmonary contusion assessments for observers A–D.

Absolute round-to-round differences were positively associated with pulmonary contusion severity for all four observers (Spearman's $\rho = 0.560$, 0.713, 0.652 and 0.656 for observers A, B, C and D, respectively; Figure 10).

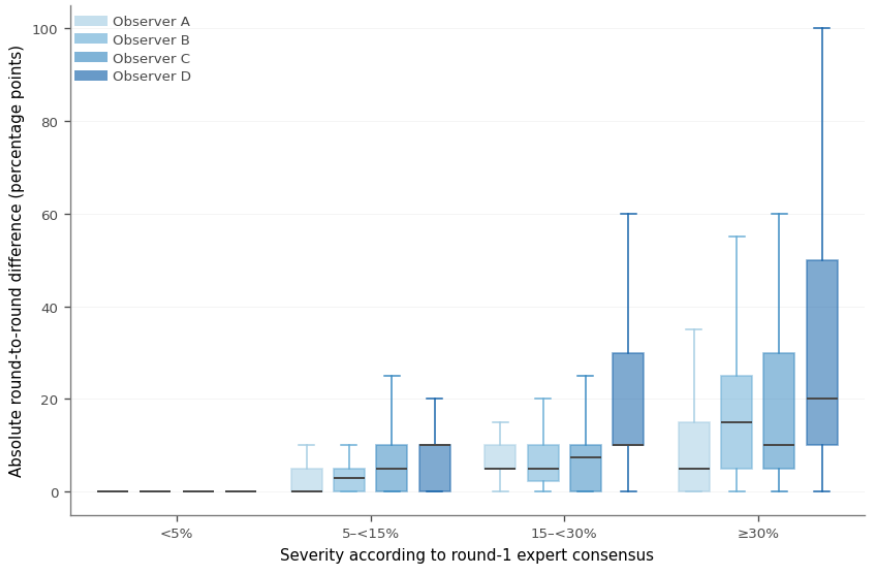


Figure 10: Absolute differences between repeated pulmonary contusion assessments across pulmonary contusion severity for observers A–D.

3.4. Pulmonary contusion quantification

3.4.1. Validation of student annotations for training-set expansion

To determine whether the limited expert-annotated dataset could be expanded, annotations from a trained Technical Medicine master's student were compared with the six-observer mean. The student independently assessed the same 47 CT scans while blinded to the clinical observer scores and their variability.

Comparability was evaluated using four predefined criteria: ICC(A,1) with the clinical observer mean, absolute mean bias, coverage of the clinical observer range, and the ratio between the student MAE and the mean leave-one-observer-out clinical observer MAE. The predefined acceptance criteria and corresponding results are presented in Table 8; the rationale for these criteria and additional details of the validation procedure are provided in Appendix D.

Table 8: Predefined acceptance criteria for validation of student annotations.

Criterion	Acceptance threshold	Result
ICC(A,1) with clinical observer mean	≥ 0.50	0.777
MAE ratio relative to clinical observers	≤ 1.10	1.006
Absolute mean bias	≤ 5.0 pp	1.96 pp
Scores within clinical observer range	$\geq 80\%$	86.9%

ICC, intraclass correlation coefficient; MAE, mean absolute error; pp, percentage points.

All four predefined acceptance criteria were satisfied. The student annotations were therefore considered suitable for training-set expansion, and the student annotated 23 additional CT scans that were included exclusively in the training portion of each cross-validation fold.

3.4.2. Loss function and batch size comparison

Four regression objectives and batch sizes were compared while keeping the network architecture, data partitions, and remaining training settings fixed. Mean squared error (MSE) served as the conventional baseline, Huber loss as a less outlier-sensitive objective, DenseWeight-weighted MSE to increase the contribution of underrepresented target values, and Balanced MSE to account for imbalance in the continuous target distribution [45]. For MSE, Huber loss, and DenseWeight-weighted MSE, batch sizes of 2, 4, and 8 were evaluated. The original MSE configuration with batch size 1 was retained as a baseline. Balanced MSE was implemented using the batch-based Monte-Carlo formulation and evaluated with batch size 8, the largest computationally feasible batch size, to provide a larger within-batch approximation of the training-label distribution [45]. DenseWeight was applied with $\alpha = 0.25$ [46].

Performance at higher pulmonary contusion severity was additionally summarised for patients with reference PC percentages above 24%, reflecting the approximate threshold previously associated with worse in-hospital outcomes [17]. Differences in patient-level absolute error between configurations were assessed using a Friedman test, accounting for repeated evaluation of the same patients across configurations.

Key cross-validation performance measures for the evaluated training configurations are summarised in Table 9; complete performance results are provided in Appendix E.

Table 9: Key cross-validation performance measures for the evaluated pulmonary contusion training configurations.

Objective	BS	Patient MAE	Segment MAE	Calibration slope	MAE $> 24\%$
MSE	1	8.69	12.85	0.31	19.09
MSE	2	6.79	11.99	0.63	10.74
MSE	4	6.46	11.60	0.59	11.80
MSE	8	6.31	11.34	0.59	12.16
Huber	2	6.77	10.19	0.60	15.22
Huber	4	6.26	10.24	0.64	13.93
Huber	8	6.41	10.14	0.57	16.33
DenseWeight-MSE	**2**	**6.19**	**11.88**	**0.71**	**8.08**
DenseWeight-MSE	4	6.40	11.77	0.67	8.64
DenseWeight-MSE	8	6.30	11.55	0.64	10.04
Balanced MSE	8	9.54	14.42	0.47	9.23

MAEs are reported in percentage points. The selected training configuration is shown in bold. Severity-specific MAE was calculated for patients with reference pulmonary contusion percentages above 24%. BS, batch size; MAE, mean absolute error; MSE, mean squared error.

Overall patient-level absolute errors did not significantly differ between training configurations (Friedman $\chi^2(10) = 15.60$, $p = 0.112$; Kendall's $W = 0.042$). DenseWeight-weighted MSE with batch size 2 achieved the lowest patient-level MAE and the highest calibration slope among the evaluated configurations. Huber loss resulted in lower segment-level MAE, whereas Balanced MSE showed the highest overall patient-level MAE.

Given the absence of a significant overall difference in patient-level absolute error, DenseWeight-weighted MSE with batch size 2 was selected based on its combination of the lowest patient-level MAE and a calibration slope closest to one.

3.4.3. Independent test-set evaluation

Independent test-set performance is summarised in Table 10. Performance was stronger at patient level than at segment level. Segment-level predictions covered a narrower range than the reference values. Figure 11 shows the patient-level predictions against the clinical consensus, with lower reference PC values tending to be overestimated and prediction errors becoming more variable at higher severity.

Table 10: Independent test-set performance of the selected pulmonary contusion model.

Metric	Patient	Segment
n	10	120
MAE (pp)	5.70	10.23
Mean bias (pp)	2.32	2.32
Pearson correlation (r)	0.904	0.654
Calibration slope	0.677	0.500
Prediction range (%)	3.0–37.8	0.0–60.1
Reference range (%)	2.2–37.6	0.0–82.5

MAE, mean absolute error; pp, percentage points.

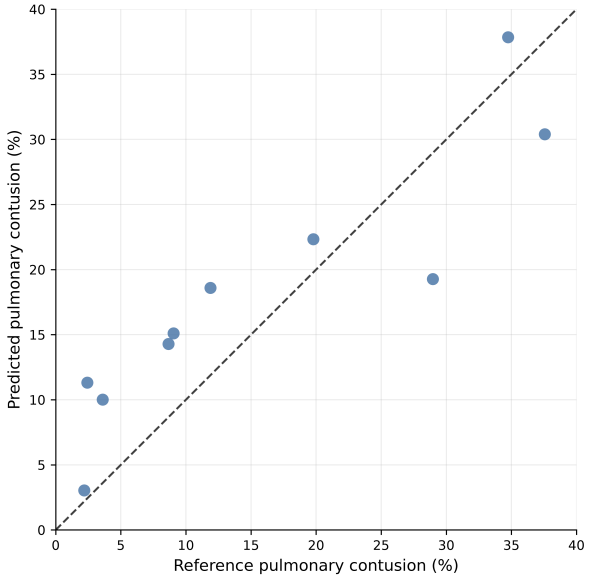


Figure 11: Reference and predicted patient-level pulmonary contusion percentages on the independent test set. The dashed line represents perfect agreement.

3.5. Pneumonia prediction model

3.5.1. Primary matched analysis

The primary analysis evaluated whether the newly developed deep learning-derived pulmonary injury measurements could replace the corresponding imaging features in the original combined pneumonia prediction model. Three models are distinguished: the clinical-only model, containing only clinical predictors; the original combined model, containing clinical and conventionally derived radiological predictors; and the deep learning (DL)-feature matched model, in which the original pulmonary injury measurements were replaced by the newly developed measurements (Table 11).

For the DL-feature matched analysis, the original modelling strategy and predefined data partitions were retained. Pneumothorax and haemothorax were represented by absolute and relative volumes, while pulmonary contusion was represented by absolute volume, relative extent and binary presence.

Across the evaluated LASSO regularisation values and classifiers, the highest validation AUC was obtained at $\alpha = 0.0125$ by both Naive Bayes and support vector machine. Naive Bayes was retained in accordance with the original modelling procedure. LASSO selected pulmonary contusion volume, pneumothorax percentage, haemothorax percentage, mechanical ventilation and parenchyma percentage.

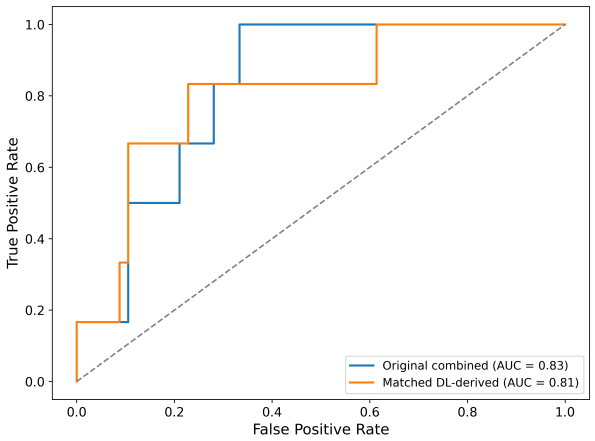


Figure 12: Receiver operating characteristic curves on the held-out test set for the original combined model and the DL-feature matched model using the newly developed deep learning-derived pulmonary injury measurements. AUC, area under the curve; DL, deep learning.

Following retraining on the combined training and validation sets, the DL-feature matched model showed similar discrimination to the original combined model on the independent test set. Both models showed higher test AUCs than the clinical-only model (Table 11 and Figure 12).

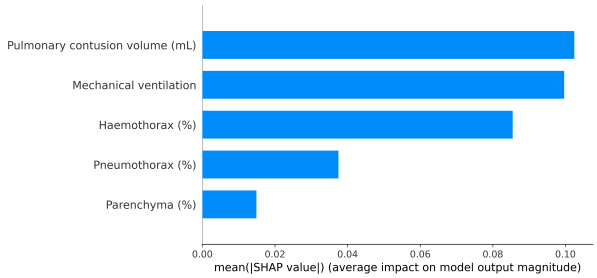
Table 11: Comparison of the clinical-only, original combined and DL-feature matched pneumonia prediction models.

Model	Selected predictors	Class.	Val. AUC	Test AUC	Sens.	Spec.
Clinical-only	Flail chest, chest tube, asthma, age, diabetes, mechanical ventilation	LR	0.81	0.74	0.75	0.67
Original combined	PC (mL), PTX (%), HTX (%), mechanical ventilation, bilateral rib fractures, age	NB	0.81	0.83	1.00	0.67
DL-feature matched	PC (mL), PTX (%), HTX (%), mechanical ventilation, parenchyma (%)	NB	0.83	0.81	0.83	0.77

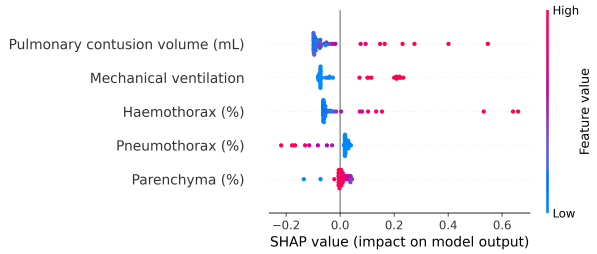
AUC, area under the curve; Class., classifier; DL, deep learning; HTX, haemothorax; LR, logistic regression; NB, Naive Bayes; PC, pulmonary contusion; PTX, pneumothorax; Sens., sensitivity; Spec., specificity; Val., validation.

SHAP analysis showed that pulmonary contusion volume had the largest mean absolute contribution to predictions in the DL-feature matched model, followed by mechanical ventilation and haemothorax percentage (Figure 13A-B). Higher pulmonary contusion volume and haemothorax percentage, as well as mechanical ventilation, were predominantly associated with increased predicted pneumonia risk, whereas higher pneumothorax percentage was predominantly associated with lower predicted risk within the fitted model. Corresponding SHAP analyses for the original clinical-only and combined clinical–radiological models are provided in Ap-

pendix F for comparison.



(a) Mean absolute SHAP values



(b) SHAP summary plot

Figure 13: Interpretation of the DL-feature matched model. (a) Mean absolute SHAP values indicating global feature importance. (b) SHAP summary plot showing the direction and magnitude of individual feature contributions. DL, deep learning; SHAP, Shapley additive explanations.

3.5.2. Exploratory optimisation

Three exploratory modifications of the imaging feature representation were evaluated using the same modelling procedure as the DL-feature matched analysis (Table 12). Parenchyma volume and percentage were first excluded because these variables do not directly quantify traumatic pulmonary injury, resulting in numerically higher validation and test AUCs. Applying post-processing thresholds for pneumothorax and haemothorax did not improve discrimination. Finally, additional descriptors of injury severity and anatomical distribution were evaluated, including severity categories, bilateral involvement and regional pulmonary contusion measures. None of these additional descriptors was selected by LASSO, and predictive performance remained unchanged.

Overall, exclusion of parenchyma was the only modification associated with numerically higher AUCs. Given the exploratory nature of these analyses and the small independent test cohort, these numerical differences were not interpreted as evidence of superior predictive performance.

Table 12: Performance of the DL-feature matched model and exploratory feature optimisation experiments.

Metric	DL-feature matched	No parenchyma	Thresholded PTX/HTX	Extended descriptors
LASSO α	0.0125	0.0125	0.0125	0.020
Classifier	NB	NB	NB	SVM
Validation AUC	0.83	0.84	0.84	0.83
Test AUC	0.81	0.82	0.81	0.81
Sensitivity	0.83	0.83	0.83	0.83
Specificity	0.77	0.68	0.68	0.77

AUC, area under the curve; DL, deep learning; HTX, haemothorax; LASSO, least absolute shrinkage and selection operator; NB, Naive Bayes; PTX, pneumothorax; SVM, support vector machine.

4

Discussion

4.1. Principal findings

This study investigated whether thoracic injury burden following blunt thoracic trauma can be automatically and reliably quantified from admission CT and whether these quantitative imaging features provide additional information for prediction of post-traumatic pneumonia. Automated quantification was feasible for pneumothorax and haemothorax, although haemothorax remained more challenging. For pulmonary contusion, substantial inter- and intraobserver variability highlighted uncertainty in visual assessment, and the mean assessment of six clinical observers was therefore used as a consensus reference for model development. Automated pulmonary contusion quantification was feasible particularly at patient level, although regional performance remained more limited. Incorporation of the newly developed pulmonary injury measurements into the existing pneumonia prediction framework resulted in similar test-set discrimination to the original combined model.

4.2. Pleural injury quantification

Automated pleural injury quantification was more accurate for pneumothorax than for haemothorax. Pneumothorax was detected with high sensitivity and showed relatively low volumetric error, whereas haemothorax segmentation showed lower spatial overlap and systematic volumetric underestimation. Previous pneumothorax segmentation studies reported Dice scores of approximately 0.87–0.96, although differences in study design limit direct comparison [24, 25, 23, 26, 27, 28]. Several studies evaluated only pneumothorax-positive cases, whereas the present pipeline was required to both detect and quantify injury in a heterogeneous trauma population. Similarly, the higher haemothorax Dice of 0.75 reported by Dreizin et al. was obtained in a population selected for haemothorax and excluding trace collections [21]. These differences underline the importance of considering detection and volumetric accuracy alongside spatial overlap when evaluating quantitative injury segmentation.

Injury volume remains relevant when interpreting quantitative performance. Small pneumothoraces and haemothoraces are relatively sensitive to segmentation inaccuracies because small absolute errors represent a larger proportion of the total injury volume, consistent with previous findings for pneumothorax quantification at low volumes [27]. All false-negative haemothoraces had non-zero predicted volumes that fell below the 40-mL post-processing threshold and were therefore classified as negative. This illustrates the trade-off introduced by post-processing: suppressing small false-positive predictions may also remove small true-positive predictions. Conversely, the largest haemothoraces were correctly detected, although substantial volumetric underestimation could still occur. Thus, balancing suppression of small false-positive predictions against retention of true small injuries, while maintaining accurate quantification of larger collections, remains an important challenge.

The ablation experiment demonstrated the value of anatomical context for haemothorax segmentation. Removing the parenchyma class significantly increased haemothorax volume error and reduced specificity, whereas pneumothorax performance changed little. Explicit modelling of the lung parenchyma therefore appears to help distinguish pleural fluid from adjacent structures and reduce false-positive haemothorax segmentation. This supports the use of a multi-class approach in the present pipeline, although separate pneumothorax and haemothorax models were not evaluated.

4.3. Pulmonary contusion assessment and quantification

Pulmonary contusion posed a different challenge from pleural injury quantification because its extent lacks a distinct anatomical boundary and therefore partly depends on visual interpretation. This was reflected by moderate agreement between individual clinical observers and substantial differences in intraobserver repeatability. Both inter- and intraobserver disagreement increased with pulmonary contusion severity. In the absence of an objective reference standard, the mean assessment of six clinical observers was used as a consensus reference for model development. Although observer disagreement remained, this approach provided a more robust reference than reliance on an individual observer.

The limited expert-annotated dataset was expanded only after student annotations satisfied all predefined comparability criteria against the clinical consensus. The additional student-annotated scans were restricted to model training, thereby increasing the available training data while retaining an expert-based standard for validation and testing. However, these additional cases were labelled by a single observer, and the acceptance criteria were pragmatic study-specific thresholds rather than externally validated criteria for annotation equivalence.

On the independent test set, pulmonary contusion quantification was more accurate at patient than segment level,

suggesting that averaging across regions may reduce the influence of local errors. This is favourable for its intended use as a measure of overall injury burden. Calibration slopes below one indicated reduced variation in predicted severity relative to the reference, particularly at segment level. Errors were more variable at higher severity, which may be relevant because greater pulmonary contusion extent has been associated with worse in-hospital outcomes [17]. This difficulty at higher severity may partly reflect the limited number of severe cases available for training, although the small independent test set limits conclusions regarding severity-dependent performance.

4.4. Pneumonia prediction

Incorporation of the newly developed quantitative pulmonary injury measurements into the existing prediction framework preserved pneumonia discrimination at a level similar to that of the original combined model. The technical performance of the original and newly developed injury measurements is not directly comparable because their target definitions and evaluation strategies differed. The original pipeline evaluated segmentation on selected two-dimensional slices and injury detection against clinical diagnoses, whereas the present models used full-volume reference segmentations for pleural injury and multi-observer regional assessments for pulmonary contusion. This illustrates that methodological improvements in upstream injury quantification do not necessarily translate into improved downstream prediction.

Pulmonary contusion volume showed the largest mean absolute SHAP contribution in the DL-feature matched model, consistent with previous literature identifying pulmonary contusion as an important risk factor for pneumonia and other pulmonary complications after thoracic trauma [6, 8]. In particular, increasing contusion extent has been associated with adverse pulmonary outcomes [17]. Mechanical ventilation and haemothorax percentage also predominantly contributed towards higher predicted pneumonia risk. In contrast, higher pneumothorax percentages predominantly contributed towards lower model predictions. This should not be interpreted as evidence of a protective effect, as SHAP reflects contributions within the fitted model rather than independent or causal associations and may reflect correlations or interactions between predictors.

Non-selection of the additional descriptors does not establish that injury distribution or bilateral involvement is clinically irrelevant, as their contribution may have been redundant with existing predictors or difficult to detect in the available sample. Interpretation of all pneumonia prediction results remains limited by the small independent test cohort, in which only six of 63 patients developed pneumonia.

4.5. Clinical implications

The potential clinical relevance of the quantitative imaging features developed in this study lies in their automated availability early after admission. In contrast to previous prediction studies in which thoracic injuries were commonly represented by binary presence or manually derived severity measures, the present workflow provides automatically derived continuous measures of pneumothorax, haemothorax and pulmonary contusion extent directly from the admission CT [32, 33, 34, 35].

Because pneumonia generally develops later during the clinical course, these measurements could contribute to risk stratification before the complication becomes clinically apparent. Their automated derivation may facilitate integration into an early prediction workflow and complement existing clinical assessment. The prominent contribution of pulmonary contusion volume within the fitted model further supports the potential value of quantitative injury extent, although this does not establish superiority over categorical injury representations or clinical utility.

4.6. Strengths and limitations

A major strength of this study was the construction and evaluation of the pulmonary contusion consensus reference using assessments from six clinical observers. This allowed observer agreement to be quantified and reduced dependence on a single subjective assessment. A further strength was evaluation in a heterogeneous trauma population including injury-positive and injury-negative cases and concurrent thoracic injuries. This reflects the intended application more closely than evaluation restricted to selected injury-positive cases.

The principal limitation was the amount of available data, particularly for pulmonary contusion quantification and pneumonia prediction. The pulmonary contusion model was independently tested in only ten patients. Additionally, only six pneumonia events occurred among the 63 patients in the pneumonia test set. Performance estimates are therefore sensitive to individual cases, limiting conclusions regarding generalisability and differences between model configurations.

Reference-standard quality differed between study components. Pulmonary contusion assessment was based on six independent clinical observers, whereas pneumothorax and haemothorax reference segmentations were delineated by a single annotator and their inter- and intraobserver variability was therefore unknown. The pneumonia outcome was derived from routinely collected clinical data, in which different pneumonia aetiologies could not always be distinguished. These sources of label uncertainty, together with errors in automatically derived rib fracture and pulmonary injury features, may affect downstream pneumonia prediction.

Finally, this was a retrospective single-centre study without external validation. Differences in patient populations, CT acquisition and clinical management may affect performance in other settings. Validation in larger, preferably prospective multicentre cohorts is therefore required before clinical applicability can be established.

4.7. Future work

Future work should first focus on further development and external validation of the automated pulmonary injury measurements in larger and more diverse datasets. This is particularly important for pulmonary contusion, for which the limited expert-annotated dataset restricted representation of the full severity range. The annotation strategy used in this study may provide a scalable approach in which multi-observer assessment of a subset is used to establish a consensus reference and evaluate additional trained annotators, whose assessments can subsequently expand the training dataset.

The complete automated workflow should subsequently be evaluated in external and preferably prospective multicentre cohorts. Such studies should use a standardised pneumonia definition to reduce outcome-label uncertainty and define a clear prediction time point, restricting predictors to information available at that time. This would establish whether predictive performance is maintained across differences in patient populations, CT acquisition protocols and clinical practice.

Finally, future research should investigate how pneumonia risk estimates could contribute to clinical decision-making rather than prediction alone. A high predicted risk does not necessarily imply benefit from a particular intervention. Randomised controlled studies could evaluate whether treatment effects differ according to predicted risk or injury profile, and ultimately determine whether incorporating automated risk estimates into clinical decision-making improves patient outcomes.

5

Conclusion

This study investigated automated quantification of pulmonary injury after blunt thoracic trauma and its integration into prediction of post-traumatic pneumonia. Pneumothorax could be quantified with high detection sensitivity and relatively low volumetric error, whereas haemothorax remained more challenging, particularly regarding volumetric underestimation. Explicit modelling of the lung parenchyma improved haemothorax quantification and reduced false-positive predictions.

Pulmonary contusion assessment showed substantial inter- and intraobserver variability, highlighting the uncertainty inherent in visual assessment. The mean assessment of six clinical observers was therefore used as a consensus reference for model development. Automated pulmonary contusion quantification was feasible, particularly at patient level, although regional performance remained more limited.

Integration of the resulting automated pulmonary injury measurements into the existing pneumonia prediction framework largely preserved discrimination compared with the original combined model. Overall, this study demonstrates that pneumothorax, haemothorax and pulmonary contusion burden can be automatically quantified from trauma CT and translated into quantitative imaging features for pneumonia risk prediction. By providing continuous measures of injury extent directly from admission imaging, this approach forms a basis for further development of automated, imaging-informed risk assessment following blunt thoracic trauma.

References

1. Bellone A, Bossi I, Etteri M, Cantaluppi F, Pina P, Guanziroli M, et al. Factors associated with ICU admission following blunt chest trauma. Can Respir J 2016; 2016:3257846. DOI: 10.1155/2016/3257846
2. Prêtre R, Chilcott M. Blunt trauma to the heart and great vessels. N Engl J Med 1997; 336(9):626–32. DOI: 10.1056/NEJM199702273360906
3. Baker E, Battle C, Lee G. Blunt mechanism chest wall injury: initial patient assessment and acute care priorities. Emerg Nurse 2024; 32(3):34–42. DOI: 10.7748/en.2024.e2181
4. Eghbalzadeh K, Sabashnikov A, Zeriouh M, Choi YH, Bunck AC, Mader N, et al. Blunt chest trauma: a clinical chameleon. Heart. 2018; 104(9):719–24. DOI: 10.1136/heartjnl-2017-312111
5. Narayanan R, Kumar S, Gupta A, Bansal VK, Sagar S, Singhal M, et al. An analysis of presentation, pattern and outcome of chest trauma patients at an urban level 1 trauma center. Indian J Surg 2018; 80(1):36–41. DOI: 10.1007/s12262-016-1554-2
6. Cohn SM. Pulmonary contusion: review of the clinical entity. J Trauma 1997; 42(5):973–9. DOI: 10.1097/00005373-199705000-00033
7. Shorr RM, Crittenden M, Indeck M, Hartunian SL, Rodriguez A. Blunt thoracic trauma: analysis of 515 patients. Ann Surg 1987; 206(2):200–5. DOI: 10.1097/00000658-198708000-00013
8. Park HO, Kang DH, Moon SH, Yang JH, Kim SH, Byun JH. Risk factors for pneumonia in ventilated trauma patients with multiple rib fractures. Korean J Thorac Cardiovasc Surg 2017; 50(5):346–54. DOI: 10.5090/kjtcs.2017.50.5.346
9. Richardson JD, Adams L, Flint LM. Selective management of flail chest and pulmonary contusion. Ann Surg 1982; 196(4):481–7. DOI: 10.1097/00000658-198210000-00012
10. Antonelli M, Luisa Moro M, Capelli O, De Blasi RA, D'Errico RR, Conti G, et al. Risk factors for early onset pneumonia in trauma patients. Chest 1994; 105(1):224–8. DOI: 10.1378/chest.105.1.224
11. Mahmood I, Tawfeek Z, El-Menyar A, Zarour A, Afifi I, Kumar S, et al. Outcome of concurrent occult hemothorax and pneumothorax in trauma patients who required assisted ventilation. Emerg Med Int 2015; 2015:859130. DOI: 10.1155/2015/859130
12. Hofman M, Andruszkow H, Kobbe P, Poeze M, Hildebrand F. Incidence of post-traumatic pneumonia in poly-traumatized patients: identifying the role of traumatic brain injury and chest trauma. Eur J Trauma Emerg Surg 2020; 46(1):11–9. DOI: 10.1007/s00068-019-01179-1
13. Battle C, Carter K, Newey L, Giamello JD, Melchio R, Hutchings H. Risk factors that predict mortality in patients with blunt chest wall trauma: an updated systematic review and meta-analysis. Emerg Med J 2023; 40(5):369–78. DOI: 10.1136/emermed-2021-212184
14. Trabelsi B, Ghorbel S, Ben Rabeh R, Bouassida M, Ben Ali M. C-reactive protein in the early diagnosis of pneumonia complicating severe blunt chest trauma. Tunis Med 2023; 101(10):756–8
15. Lafferty PM, Anavian J, Will RE, Cole PA. Operative treatment of chest wall injuries: indications, technique, and outcomes. J Bone Joint Surg Am 2011; 93(1):97–110. DOI: 10.2106/JBJS.I.00696
16. Sharma VJ, Summerhayes R, Wang Y, Kure C, Marasco SF. Surgical stabilisation of rib fractures: a meta-analysis of randomised controlled trials. Injury 2024; 55(8):111705. DOI: 10.1016/j.injury.2024.111705
17. Van Diepen MR, Wijffels MME, Verhofstad MHJ, Van Lieshout EMM. Classification methods of pulmonary contusion based on chest CT and the association with in-hospital outcomes: a systematic review of literature. Eur J Trauma Emerg Surg 2024; 50(6):2727–40. DOI: 10.1007/s00068-024-02666-w
18. Bilello JF, Davis JW, Lemaster DM. Occult traumatic hemothorax: when can sleeping dogs lie? Am J Surg 2005; 190(6):844–8. DOI: 10.1016/j.amjsurg.2005.05.053
19. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Semin Radiat Oncol 2019; 29(3):185–97. DOI: 10.1016/j.semradonc.2019.02.001
20. Cai W, Tabbara M, Takata N, Yoshida H, Harris GJ, Novelline RA, et al. MDCT for automated detection and measurement of pneumothoraces in trauma patients. AJR Am J Roentgenol. 2009; 192(3):830–6. DOI: 10.2214/ajr.08.1339
21. Dreizin D, Nixon B, Hu J, Albert B, Yan C, Yang G, et al. A pilot study of deep learning-based CT volumetry for traumatic hemothorax. Emerg Radiol 2022; 29(6):995–1002. DOI: 10.1007/s10140-022-02087-5

22. Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol 2019; 29(3):1391–9. DOI: 10.1007/s00330-018-5695-5

23. Xue B, Liu ZQ, Wang QF, Tang Q, Huang J, Zhou Y. SNU-Net: a self-supervised deep learning method for pneumothorax segmentation on chest CT. *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2022 :60–4. DOI: 10.1109/ISCAS48785.2022.9937654

24. Wu W, Liu G, Liang K, Zhou H. Pneumothorax segmentation in routine computed tomography based on deep neural networks. *2021 4th International Conference on Intelligent Autonomous Systems (ICoIAS)*. Shanghai, China: IEEE, 2021 :78–83. DOI: 10.1109/ICoIAS53694.2021.00022

25. Liu Y, Liang P, Liang K, Chang Q. Automatic and efficient pneumothorax segmentation from CT images using EFA-Net with feature alignment function. Sci Rep 2023; 13(1):15291. DOI: 10.1038/s41598-023-42388-4

26. Tang Z, Zhang J, Bai C, Zhang Y, Liang K, Yao X. Dense Swin-Unet: dense Swin Transformers for semantic segmentation of pneumothorax in CT images. J Mech Med Biol 2023; 23(8):2340069. DOI: 10.1142/S0219519423400699

27. Röhrich S, Schlegl T, Bardach C, Prosch H, Langs G. Deep learning detection and quantification of pneumothorax in heterogeneous routine chest computed tomography. Eur Radiol Exp 2020; 4(1):26. DOI: 10.1186/s41747-020-00152-7

28. Gencer A, Toker Yİ. Segmentation of pneumothorax on chest CTs using deep learning based on Unet-Resnet-50 convolutional neural network structure. Eur J Ther 2024; 30(3):249–57. DOI: 10.58600/eurjther2018

29. Daly M, Miller PR, Carr JJ, Gayzik FS, Hoth JJ, Meredith JW, et al. Traumatic pulmonary pathology measured with computed tomography and a semiautomated analytic method. Clin Imaging 2008; 32(5):346–54. DOI: 10.1016/j.clinimag.2008.02.026

30. Sarkar N, Zhang L, Campbell P, Liang Y, Li G, Khedr M, et al. Pulmonary contusion: automated deep learning-based quantitative visualization. Emerg Radiol 2023; 30(4):435–41. DOI: 10.1007/s10140-023-02149-2

31. Choi J, Mavrommati K, Li NY, Patil A, Chen K, Hindin DI, et al. Scalable deep learning algorithm to compute percent pulmonary contusion among patients with rib fractures. J Trauma Acute Care Surg 2022; 93(4):461–6. DOI: 10.1097/TA.0000000000003619

32. Seok J, Lee J, Kang WS. Artificial intelligence models for predicting pulmonary complications in patients with chest trauma: a retrospective study. J Trauma Inj 2025; 38(3):237–47. DOI: 10.20408/jti.2025.0100

33. Seok J, Yoon SY, Lee JY, Kim S, Cho H, Kang WS. Novel nomogram for predicting pulmonary complications in patients with blunt chest trauma with rib fractures: a retrospective cohort study. Sci Rep 2023; 13:9448. DOI: 10.1038/s41598-023-36679-z

34. Kobes T, Sweet AAR, IJpma FFA, Leenen LPH, Houwert RM, Wessem KJP van, et al. Identifying predictors of nosocomial pneumonia in trauma patients admitted to a level-1 trauma center. Arch Orthop Trauma Surg 2024; 145(1):100. DOI: 10.1007/s00402-024-05672-0

35. Wang L, Zhao Y, Wu W, He W, Yang Y, Wang D, et al. Development and validation of a pulmonary complications prediction model based on the Yang's index. J Thorac Dis 2023; 15(4):2213–23. DOI: 10.21037/jtd-23-378

36. Marting V, Borren N, Diepen MR van, Lieshout EMM van, Wijffels MME, Walsum T van. A deep learning-based approach to automated rib fracture detection and CWIS classification. Int J Comput Assist Radiol Surg 2025; 20(7):1381–9. DOI: 10.1007/s11548-025-03390-5

37. Bakker G. Prediction of pneumonia following traumatic rib fractures: integrating clinical data and automated CT analysis into a machine learning model. MA thesis. Delft: Delft University of Technology, 2025

38. Prins JTH, Van Lieshout EMM, Van Wijck SFM, Scholte NTB, Den Uil CA, Vermeulen J, et al. Chest wall injuries due to cardiopulmonary resuscitation and the effect on in-hospital outcomes in survivors of out-of-hospital cardiac arrest. J Trauma Acute Care Surg 2021; 91(6):966–75. DOI: 10.1097/TA.0000000000003379

39. Prins JTH, Van Lieshout EMM, Overtoom HCG, Tekin YS, Verhofstad MHJ, Wijffels MME. Long-term pulmonary function, thoracic pain, and quality of life in patients with one or more rib fractures. J Trauma Acute Care Surg 2021; 91(6):923–31. DOI: 10.1097/TA.0000000000003378

40. Centers for Disease Control and Prevention. Pneumonia (Ventilator-Associated [VAP] and Non-Ventilator-Associated Pneumonia [PNEU]) Event. Technical report. Centers for Disease Control and Prevention, 2018

41. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol Artif Intell 2023; 5(5):e230024. DOI: 10.1148/ryai.230024

42. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021; 18(2):203–11. DOI: 10.1038/s41592-020-01008-z

43. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016; 15(2):155–63. DOI: 10.1016/j.jcm.2016.02.012

44. Lyu WH, Xia F, Zhou CS, Huang M, Ding WW, Zhang S, et al. [Application of deep learning-based chest CT auxiliary diagnosis system in emergency trauma patients]. Zhonghua Yi Xue Za Zhi 2021; 101(7):481–6. DOI: 10.3760/cma.j.cn112137-20201117-03123

45. Ren J, Zhang M, Yu C, Liu Z. Balanced MSE for imbalanced visual regression. 2022. DOI: 10.48550/arXiv.2203.16427. arXiv: 2203.16427 [cs.CV]

46. Steininger M, Kobs K, Davidson P, Krause A, Hotho A. Density-based weighting for imbalanced regression. Mach Learn 2021; 110:2187–211. DOI: 10.1007/s10994-021-06023-5

A

Pleural injury model selection and post-processing

Table 14 presents the ten eligible candidate pipelines with the lowest combined PTX and HTX mean absolute volume error during cross-validation.

Table 13: Post-processing thresholds evaluated during pleural injury model selection.

Injury	Evaluated thresholds (mL)
Pneumothorax	0, 0.1, 0.25, 0.5, 1, 1.5, 2, 3, 5, 10
Haemothorax	0, 5, 10, 15, 20, 30, 40, 50, 75, 100

Table 14: Ten eligible candidate pipelines with the lowest combined PTX and HTX mean absolute error, with the best-performing candidate from the remaining configurations shown for comparison.

		Threshold (mL)		Sensitivity		MAE (mL)		
Rank	**Configuration**	**PTX**	**HTX**	**PTX**	**HTX**	**PTX**	**HTX**	**Combined**
1	**3D full-resolution**	**3**	**40**	**0.986**	**0.910**	**8.80**	**23.83**	**16.32**
2	3D full-resolution	2	40	0.986	0.910	8.83	23.83	16.33
3	3D full-resolution	1.5	40	0.986	0.910	8.85	23.83	16.34
4	3D full-resolution	1	40	0.986	0.910	8.88	23.83	16.36
5	3D full-resolution	0.5	40	0.986	0.910	8.90	23.83	16.37
6	3D full-resolution	0.25	40	0.986	0.910	8.92	23.83	16.38
7	3D full-resolution	0.1	40	1.000	0.910	8.93	23.83	16.38
8	3D full-resolution	0	40	1.000	0.910	8.94	23.83	16.39
9	3D full-resolution	3	30	0.986	0.955	8.80	24.85	16.83
10	3D full-resolution	2	30	0.986	0.955	8.83	24.85	16.84
–	2D + 3D ensemble	1	20	0.986	0.955	6.95	35.20	21.07
–	2D	0.5	10	0.971	0.910	7.44	44.46	25.95

The selected pipeline is shown in bold. The unranked rows show the eligible pipeline with the lowest combined MAE for each of the two remaining configurations. HTX, haemothorax; MAE, mean absolute error; PTX, pneumothorax.

B

Pulmonary contusion assessment protocol

The following sections reproduce the pulmonary contusion assessment guidance provided to the clinical observers. Software-specific instructions were omitted.

B.1. Assessment purpose

Observers assessed 47 CT scans for the extent of pulmonary contusion. Each lung was divided into six predefined regions, resulting in twelve regions per patient. For each region, observers assigned a pulmonary contusion percentage ranging from 0% to 100%.

B.2. Definition and interpretation

Pulmonary contusion was defined on CT as poorly demarcated areas of increased pulmonary density, presenting as ground-glass opacity or consolidation, generally without clear volume loss and often with a patchy distribution.

Observers were instructed to score abnormalities that they would clinically interpret as traumatic pulmonary contusion. Similar parenchymal abnormalities, such as atelectatic changes, could also be encountered. Observers were therefore instructed to apply their clinical interpretation consistently across all patients.

B.3. Assessment of contusion extent

For each of the twelve lung regions, observers estimated the percentage of the region affected by pulmonary contusion. The percentage represented the spatial extent of the abnormality throughout the region rather than its intensity on an individual CT slice.

Observers were instructed to scroll through the complete region in at least the axial plane before assigning a score. Coronal and sagittal reconstructions could additionally be used when considered helpful.

B.4. Percentage scoring guidance

Pulmonary contusion extent was recorded on a continuous scale from 0% to 100%. The following values were provided as guidance for interpretation of the scale:

- 0%: no pulmonary contusion;
- 10–20%: small areas of pulmonary contusion;
- 25–40%: pulmonary contusion clearly present, involving less than half of the region;
- 50–75%: a large proportion of the region affected;
- $>$ 75%: almost the entire region affected.

These values served only as guidance and were not predefined scoring categories. Observers could assign any percentage between 0% and 100%.

B.5. Assessment instructions

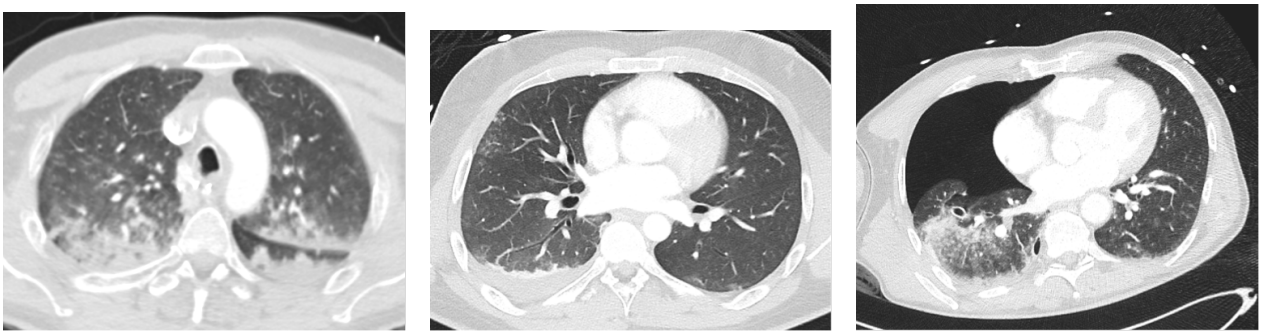
Observers were instructed to:

- scroll through the complete extent of each region before assigning a score;
- assess each region independently;
- use rounded percentage estimates, generally in increments of approximately 5–10 percentage points;

- avoid unnecessary numerical precision, for example preferring 40% or 45% over 43%;
- apply a consistent interpretation and scoring scale throughout the complete assessment set; and
- consider that the assessment set could contain CT scans without pulmonary contusion.

B.6. Reference examples

The following CT examples were provided to the clinical observers as visual guidance for the interpretation of pulmonary contusion.



(a) (b) (c)

Figure 14: Reference CT examples of pulmonary contusion provided to the clinical observers as part of the pulmonary contusion assessment instructions.

C

Detailed pulmonary contusion observer analysis

This appendix provides additional methodological detail and descriptive results for the exploratory analyses of observer disagreement across pulmonary contusion severity.

C.1. Interobserver disagreement across severity

For the analysis of interobserver disagreement across pulmonary contusion severity, each observer was compared with a leave-one-observer-out consensus. For observer j, the leave-one-observer-out consensus for segment i was calculated as

$$C_{i,-j} = \frac{\sum_{k \neq j} x_{ik}}{5},$$

where x_{ik} denotes the pulmonary contusion percentage assigned to segment i by observer k, and $C_{i,-j}$ denotes the corresponding consensus score excluding observer j.

Absolute disagreement D_{ij} was defined as

$$D_{ij} = |x_{ij} - C_{i,-j}|\,.$$

Segments were grouped according to their leave-one-observer-out consensus score into four severity categories: 0%, >0–$<10\%$, 10–$<30\%$ and $\geq 30\%$.

For descriptive comparison across severity, segment-level absolute disagreements were averaged within each patient, observer and severity category. Each patient–observer combination therefore contributed at most one value per severity category. The resulting values were summarised using the median and interquartile range (Table 15).

Table 15: Interobserver disagreement across pulmonary contusion severity categories.

Severity category	n	Median absolute disagreement [IQR] (pp)
0%	182	0.00 [0.00–0.00]
>0–$<10\%$	254	3.07 [2.13–4.68]
10–$<30\%$	221	8.33 [4.67–14.00]
$\geq 30\%$	148	17.71 [9.62–26.98]

IQR, interquartile range; pp, percentage points. n represents patient–observer combinations contributing to each severity category.

C.2. Intraobserver disagreement across severity

For the intraobserver severity analysis, pulmonary contusion severity was defined using the mean round-1 score of all six clinical observers for each segment:

$$C_i^{(1)} = \frac{1}{6}\sum_{j=1}^{6} x_{ij}^{(1)},$$

where $x_{ij}^{(1)}$ denotes the round-1 pulmonary contusion percentage assigned to segment i by observer j, and $C_i^{(1)}$ denotes the corresponding six-observer round-1 consensus score.

Absolute round-to-round disagreement for each repeated observer was calculated as

$$D_i = \left| x_i^{(2)} - x_i^{(1)} \right|,$$

where $x_i^{(1)}$ and $x_i^{(2)}$ denote that observer's first- and second-round scores for segment i, respectively.

For descriptive visualisation, segments were grouped according to the six-observer round-1 consensus into four severity categories: $< 5\%$, 5–$< 15\%$, 15–$< 30\%$ and $\geq 30\%$.

The association between pulmonary contusion severity and absolute round-to-round disagreement was additionally quantified separately for each repeated observer using Spearman's rank correlation coefficient. The correlation was calculated at the segment level using the continuous six-observer round-1 consensus score rather than the categorical severity groups.

D

Validation of student pulmonary contusion annotations

This appendix provides additional methodological detail on the validation of the student pulmonary contusion assessments against the assessments of the six clinical observers.

D.1. Reference and comparison measures

For each segment i, the clinical consensus score was defined as the mean pulmonary contusion percentage assigned by the six clinical observers:

$$C_i = \frac{1}{6}\sum_{j=1}^{6} x_{ij},$$

where x_{ij} denotes the pulmonary contusion percentage assigned to segment i by clinical observer j, and C_i denotes the corresponding six-observer clinical consensus score.

To estimate the disagreement expected for an individual clinical observer, each observer was additionally compared with a leave-one-observer-out consensus. For observer j, this consensus was calculated as

$$C_{i,-j} = \frac{\sum_{k\neq j} x_{ik}}{5},$$

where $C_{i,-j}$ denotes the consensus score for segment i after excluding observer j.

The mean absolute error for clinical observer j relative to the corresponding leave-one-observer-out consensus was calculated as

$$\mathrm{MAE}_j = \frac{1}{n}\sum_{i=1}^{n} \left| x_{ij} - C_{i,-j} \right|.$$

The clinical observer benchmark was subsequently defined as the mean of the six individual leave-one-observer-out MAEs:

$$\mathrm{MAE}_{\text{clinical}} = \frac{1}{6}\sum_{j=1}^{6} \mathrm{MAE}_j.$$

The student MAE was calculated relative to the six-observer clinical consensus:

$$\mathrm{MAE}_{\text{student}} = \frac{1}{n}\sum_{i=1}^{n} \left| s_i - C_i \right|,$$

where s_i denotes the pulmonary contusion percentage assigned by the student to segment i.

The relative magnitude of student disagreement was expressed as

$$\mathrm{MAE}_{\text{ratio}} = \frac{\mathrm{MAE}_{\text{student}}}{\mathrm{MAE}_{\text{clinical}}}.$$

A ratio of 1 therefore indicated that the student's mean absolute disagreement from the clinical consensus was equal to the mean disagreement observed for an individual clinical observer relative to the remaining clinical observers.

Table 16: Mean absolute disagreement of the student and individual clinical observers relative to their respective consensus scores.

Assessment	MAE (pp)
Clinical observer 1	5.77
Clinical observer 2	10.23
Clinical observer 3	8.12
Clinical observer 4	7.69
Clinical observer 5	12.67
Clinical observer 6	5.68
Mean clinical observer	8.36
Student	8.41

MAE, mean absolute error; pp, percentage points. For each clinical observer, MAE was calculated relative to the mean assessment of the remaining five clinical observers. Student MAE was calculated relative to the mean assessment of all six clinical observers.

The mean leave-one-observer-out MAE across the six clinical observers was 8.36 percentage points, compared with 8.41 percentage points for the student, resulting in an MAE ratio of 1.006.

D.2. Bias and observer-range coverage

Mean bias of the student assessments relative to the clinical consensus was defined as

$$\text{Bias}_{\text{student}} = \frac{1}{n}\sum_{i=1}^{n}(s_i - C_i)\,.$$

For validation, the absolute value of this mean bias was used, such that the criterion reflected the magnitude rather than the direction of systematic disagreement.

Observer-range coverage quantified the proportion of student assessments that fell within the range of scores assigned by the six clinical observers. For each segment, the clinical observer range was defined as

$$\left[\min_j(x_{ij}), \max_j(x_{ij})\right].$$

Observer-range coverage was calculated as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\min_j(x_{ij}) \le s_i \le \max_j(x_{ij})\right] \times 100\%,$$

where $\mathbb{I}$ denotes the indicator function.

D.3. Rationale for the predefined acceptance criteria

The validation criteria were defined a priori to evaluate complementary aspects of comparability between the student and clinical observer assessments. The ICC(A,1) threshold of 0.50 corresponded to the lower bound for moderate agreement according to Koo and Li [43]. The MAE-ratio threshold of 1.10 was chosen to require the student's absolute disagreement to remain within 10% of the mean disagreement observed for an individual clinical observer.

The thresholds for absolute mean bias and observer-range coverage were pragmatic study-specific criteria. An absolute mean bias threshold of 5 percentage points was used to limit systematic deviation of the student assessments from the clinical consensus. Observer-range coverage was included as a complementary measure of whether individual student assessments remained within the variability observed among the six clinical observers, with a threshold of 80% requiring this for the large majority of assessments. These thresholds were not derived from established external validation standards.

E

Detailed pulmonary contusion training-configuration performance

This appendix provides complete cross-validation results for the training configurations evaluated during pulmonary contusion model development and additional performance metrics for the final model on the independent test set.

E.1. Training-configuration performance

Complete cross-validation performance results for the evaluated loss functions and batch sizes are presented in Table 17.

Table 17: Cross-validation performance of the evaluated pulmonary contusion training configurations. Error and bias metrics are reported in percentage points. Severity-specific MAEs were calculated according to the patient-level reference pulmonary contusion percentage.

Objective	BS	Patient MAE	Patient RMSE	Segment MAE	Segment RMSE	Pearson r	Slope	Intercept	Bias	SD ratio	Prediction min	Prediction max	MAE $< 18\%$	MAE 18–24%	MAE $> 24\%$	Bias $> 24\%$
MSE	1	8.69	11.43	12.85	18.04	0.651	0.313	9.06	-1.66	0.480	1.10	35.01	5.88	5.54	19.09	-19.09
MSE	2	6.79	8.07	11.99	16.28	0.841	0.630	7.03	1.27	0.749	0.00	51.85	6.19	3.37	10.74	-10.74
MSE	4	6.46	7.90	11.60	16.14	0.858	0.590	7.08	0.69	0.688	0.27	45.20	5.48	2.66	11.80	-11.80
MSE	8	6.31	7.75	11.34	16.20	0.864	0.594	6.48	0.16	0.688	0.06	47.35	5.04	3.09	12.16	-12.16
Huber	2	6.77	8.77	10.19	16.33	0.835	0.602	3.01	-3.20	0.721	0.95	49.46	3.82	7.40	15.22	-15.22
Huber	4	6.26	8.42	10.24	16.29	0.844	0.638	2.70	-2.94	0.756	0.29	52.60	3.48	7.34	13.93	-13.93
Huber	8	6.41	8.78	10.14	16.36	0.864	0.572	2.82	-3.86	0.662	0.29	44.74	3.16	6.15	16.33	-16.33
DenseWeight	**2**	**6.19**	**7.75**	**11.88**	**16.47**	**0.859**	**0.712**	**6.59**	**2.09**	**0.829**	**0.78**	**56.87**	**6.11**	**3.52**	**8.08**	**-8.08**
DenseWeight	4	6.40	7.65	11.77	16.26	0.869	0.669	7.32	2.15	0.770	0.24	52.20	6.19	3.79	8.64	-8.64
DenseWeight	8	6.30	7.58	11.55	16.11	0.869	0.639	7.01	1.39	0.736	0.10	50.07	5.69	3.27	10.04	-10.04
Balanced MSE	8	9.54	11.29	14.42	18.21	0.741	0.471	13.71	5.45	0.635	0.84	46.69	10.54	5.24	9.23	-9.23

BS, batch size; MAE, mean absolute error; MSE, mean squared error; RMSE, root mean squared error; SD, standard deviation. The selected training configuration is shown in bold.

E.2. Independent test-set performance

Additional segment- and patient-level performance metrics for the final pulmonary contusion model on the independent test set are presented in Table 18.

Table 18: Detailed performance of the final pulmonary contusion model on the independent test set. Error and bias metrics are reported in percentage points.

Metric	Segment level	Patient level
n	120	10
MAE (pp)	10.23	5.70
RMSE (pp)	14.27	6.29
Mean bias (pp)	2.32	2.32
Pearson r	0.654	0.904
Spearman ρ	0.729	0.964
Calibration slope	0.500	0.677
Calibration intercept (pp)	10.27	7.46
SD ratio	0.764	0.749
Prediction range (%)	0.0–60.1	3.0–37.8

MAE, mean absolute error; pp, percentage points; RMSE, root mean squared error; SD, standard deviation. At segment level, n represents lung regions; at patient level, n represents patients.

F

Original pneumonia prediction model feature importance

For comparison with the matched pneumonia prediction model presented in the main results, SHAP-based feature importance for the original clinical-only and combined clinical–radiological models is shown in Figure 15.

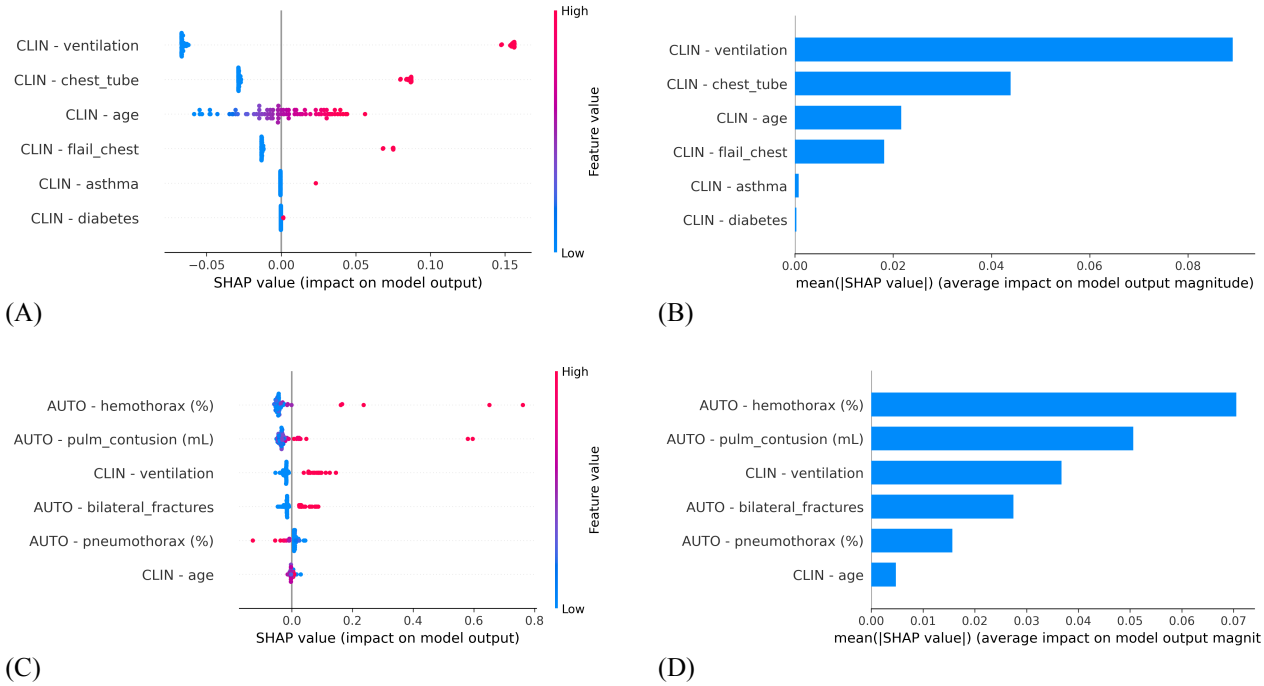


Figure 15: SHAP-based feature importance for the original pneumonia prediction models. (A–B) Clinical-only model. (C–D) Combined clinical–radiological model.