Deep Learning-Based Magnetic Resonance Imaging Lung Segmentation and Volumetric Marker Extraction in Preterm Infants

Authors

Benedikt Mairhörmann[1,§], Alejandra Castelblanco[1,§], Friederike Häfner[2,4,§], Vanessa Pfahler[5], Lena Haist[2], Dominik Waibel[1], Andreas Flemmer[4], Harald Ehrhardt[7], Sophia Stoecklein[5], Olaf Dietrich[5], Kai Foerster[4], Anne Hilgendorff[2,3,†], Benjamin Schubert[1,6, †,*]

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

[2]Institute for Lung Biology and Disease and Comprehensive Pneumology Center, Helmholtz Zentrum München, Germany, Member of the German Center for Lung Research (DZL)

[3]Center for Comprehensive Developmental Care (CDeC[LMU]) Hospital of the Ludwig-Maximilians University, Munich, Germany

[4]Department of Neonatology, Perinatal Center, Hospital of the Ludwig-Maximilians University, Munich, Germany

[5]Department of Radiology, Hospital of the Ludwig-Maximilians University, Munich, Germany

[6]Department of Mathematics, Technical University of Munich, 85748 Garching bei München, Germany

[7]Department of General Pediatrics & Neonatology, Justus-Liebig-University, Giessen, Germany, Member of the German Center for Lung Research (DZL)

[§]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

† The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

*Corresponding author: benjamin.schubert@helmholtz-muenchen.de

Abstract

Objective and Impact Statement: We apply a deep learning (DL) segmentation method and automate the extraction of imaging markers for neonatal lung structure using magnetic resonance imaging (MRI) in order to inform clinical care with robust and quantifiable information about the neonatal lung.

Introduction: Quantification of lung structural information in a standardized fashion is crucial to inform diagnostic processes that enable personalized treatment and monitoring strategies. Increased efficiency and accuracy in image quantification is especially needed in prematurely born infants, for whom long-term survival is critically determined by acute and chronic pulmonary complications, currently diagnosed based on clinical criteria due to the lack of routinely applicable diagnostic tools.

Methods: We prospectively enrolled 107 premature infants in two clinical centers with and without chronic lung disease, i.e., Bronchopulmonary Dysplasia (BPD) to perform quiet-breathing lung MRI. An ensemble of deep convolutional neural networks was developed to perform lung segmentation, with a subsequent reconstruction of the 3-dimensional lung and computation of MRI volumetric measurements and compared to the standard manual segmentation.

Results: The DL model successfully annotates lung segments with a volumetric dice score of 0.908 (Site 1) and 0.880 (Site 2), thereby reaching expert-level performance while demonstrating high transferability between study sites and robustness towards technical (low spatial resolution, movement artifacts) and disease conditions. Estimated lung volumes correlated with infant lung function testing measures and enabled the separation of neonates with and without BPD.

Conclusion: Our work demonstrates the potential of AI-supported MRI measures to perform monitoring of neonatal lung development and characterization of respiratory diseases in this high-risk patient cohort.

Keywords

Bronchopulmonary Dysplasia, Chronic Lung Disease, Preterm Infant, Lung Segmentation, Lung Magnetic Resonance Imaging, BPD Severity Prediction, Deep Learning.

MAIN TEXT

1. Introduction

Clinical decision-making in lung disease is mostly based on clinical observations and lung function measurements, ideally complemented by structural information from imaging strategies. Although quantification of lung structural information in a standardized fashion can critically inform the diagnostic process and help to implement personalized treatment and monitoring strategies, the much needed increase in efficiency, accuracy and comparability in image quantification most often remains an unmastered challenge in critical areas of today's clinical care.

These unmet diagnostic clinical needs become especially apparent in the most vulnerable, yet diagnostically challenging patient cohort. Born extremely immature, the preterm infant postnatally faces the inevitable development of acute lung injury, subsequently evolving into a chronic condition in the majority of cases [1]. While the diagnostic process still solely relies on clinical indicators of late-stage pulmonary function [2,3], the standardized assessment of radiation-free, sensitive imaging strategies would allow for diagnosing and monitoring respiratory disease from neonatal life into adulthood.

Given the low sensitivity and diagnostic value of conventional chest radiography to sensitively discriminate disease conditions and the limitations of Computed Tomography (CT) due to radiation exposure [4,5], alternative imaging techniques such as Magnetic Resonance Imaging (MRI) are being explored to provide quantitative information with prognostic relevance when assessing the diseased neonatal lung [6,7,8]. MRI in the neonatal lung is technically challenged by small subject sizes, lower spatial resolution, and sensitivity to infant motion, resulting in blurring, ghosting, and other image artifacts [9]. These conditions demand expert knowledge to obtain measurements from the acquired pulmonary images [7,10] and affect inter-rater concordances resulting in low standardization and limitiations in high-throughput of MRI-based monitoring in neonatal lung disease.

We therefore developed a deep learning-based system to support robust and standardized MRI analysis in the lungs of preterm neonates. To this end, we prospectively enrolled 107 cases with and without chronic lung disease, also known as Bronchopulmonary Dysplasia (BPD), from two study sites undergoing quiet-breathing lung MRI near term age. BPD is a chronic lung injury

syndrome resulting from a structurally immature lung condition that comprises a range of functional abnormalities (e.g., alveolar septation, airway injuries, decreased microvascular development) leading to insufficient gas exchange, with severity of the disease determined by the need of supplemental oxygen [1]. We combined recent advances in computational methods, i.e., convolutional neural networks (CNN), which have been integrated in various biomedical imaging applications [11,12,13], to test the applicability and robustness of deep learning (DL) methods for performing MRI lung segmentation in preterm infants with and without BPD. Subsequently, the obtained lung segmentations were used to compute MRI-based 3-dimensional (3D) lung volumetric features, that enabled the accurate separation of healthy cases from premature infants with different BPD severity grades.

2. Results

The proposed automated pipeline was developed and validated in a cohort of 107 preterm infants near term during quiet sleep. Quiet-breathing T2-weighted single-shot fast-spin-echo 3T-MRI lung sequences and clinical information were acquired at two study sites (Fig. 1A, Table 1).

We trained a set of U-Net CNN models to perform 2D lung segmentations on each slice of the collected neonatal MRI scans, with each model based on manual annotations of three different physicians (Fig. 1B-C) and combined them through pixel-wise majority voting (MV) to an ensemble model (Fig. 1D). A 3D representation of the lung and a clustering method to separate left and right lobes were used to calculate volumetric features (Fig. 1E).

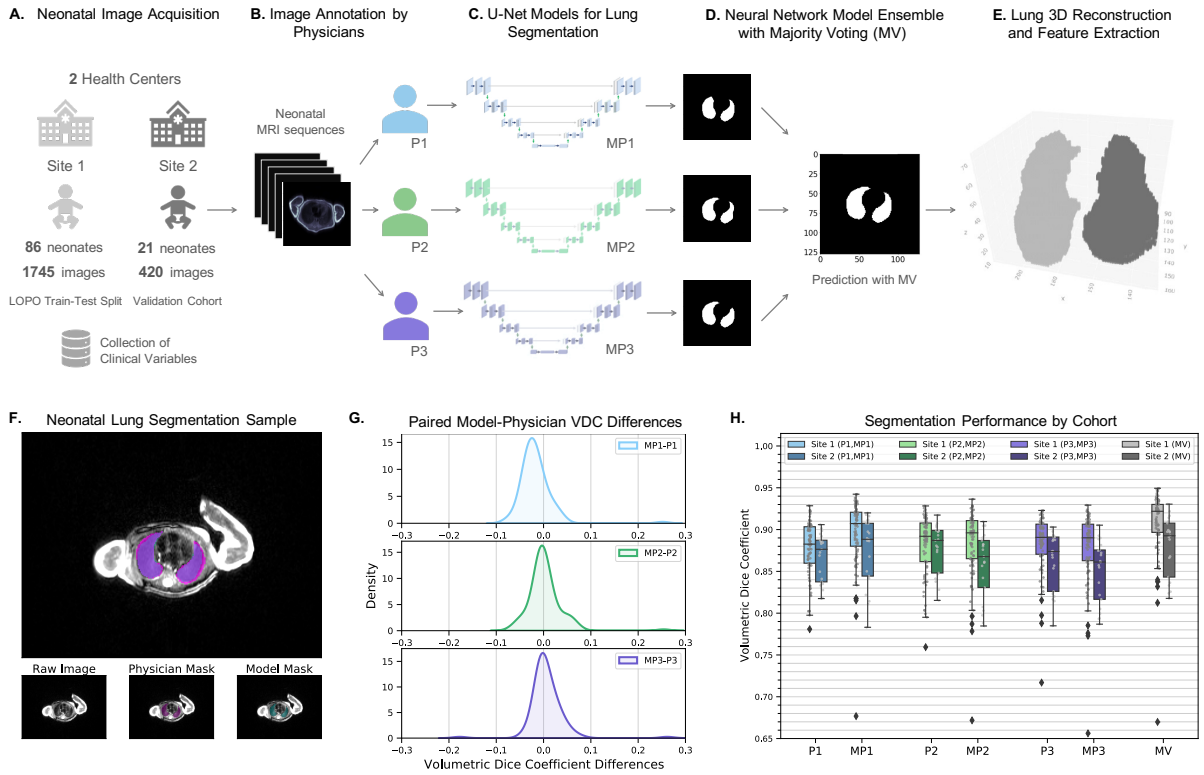


Figure 1. MRI-based Neonatal Lung Volume Analysis Pipeline. (**A**) MRI neonatal image acquisition and data collection. (**B**) Manual image annotation performed by three trained physicians. (**C-D**) CNN model training and model prediction with majority voting. (**E**) Lung 3D reconstruction and volume-based feature extraction. (**F**) Lung segmentation example. (**G**) Paired Volumetric Dice Coefficient (VDC) differences between the reference CNN model and the corresponding manual annotation. (**H**) Manual annotation and model performances by cohort.

2.1 Deep Learning Enables Human-Level Neonatal Lung Segmentation in MRI

To investigate whether human-level performance could be achieved, the segmentations generated by one of the DL models (e.g., MP1), which was trained with the manual annotations from one physician (e.g., P1), were compared with each ground-truth of the remaining manual segmentations (i.e., P2 and P3), we report the average model performance across these comparisons. As evaluation metric, we used the volumetric dice coefficient (VDC) integrating the segmentation performance of all pixels and slices from one MRI sequence. The average inter-rater VDC concordance between physicians (e.g., P1 vs each of the manual annotations P2 and P3) was also calculated as a reference for the model performance.

For study Site 1, models were trained in a leave-one-patient-out (LOPO) cross-validation scheme to generate unbiased performance estimates, while samples from study Site 2 were solely used for validation, performance for both cohorts was analyzed.

The proposed DL models achieved high segmentation performances comparable to the reference manual performance, while demonstrating low average VDC differences between the model and the reference manual segmentation when comparing identical MRI sequences (-0.017±0.035, 0.003±0.036, 0.006±0.037, for MP1, MP2, MP3; Fig. 1G). One case even exceeded average performance (one-sided Wilcoxon signed-rank test, p-value=0.0565 for MP3). Small differences between the paired performance of the model and the manual segmentation indicates an adequate representational power of the proposed models to abstract the segmentation knowledge from the provided training dataset.

The average VDC of the manual annotations were 0.875±0.032, 0.881±0.034, and 0.879±0.035 (P1, P2, P3), whereas the DL models had VDCs of 0.89±0.041, 0.878±0.042, and 0.872±0.043 (MP1, MP2, MP3), demonstrating differences of less than 0.016 points in the overall performance thus indicating high comparability between models and manual annotations.

The MV ensemble model prediction was evaluated by comparing its performance against a ground truth generated with all manual annotations (P1, P2, P3) also aggregated through pixel-wise majority voting. The MV ensemble model showed the highest segmentation performance (Fig. 1H), with an average VDC of 0.902 ±0.039 improving by 2.1% over the highest manual segmentation's VDC average, indicating human-level accuracy of the AI based segmentation method for quiet-breathing neonatal lung MRI.

2.2 Robust Automatic Segmentations Across Clinical Sites and Diseased Lungs

Performance across different study sites was assessed to confirm the models generalizability (Fig. 1H). The average VDC of the MV ensemble model was 0.908±0.039 for study Site 1 and 0.880±0.036 for study Site 2. Differences in the average VDC between cohorts were below 0.0286 points for all trained models, indicating transferability of the models between cohorts while facing minor changes in the segmentation performance. Minor differences in performance across sites were also observed in the manual segmentations, potentially originating from imbalances in image

quality demonstrated by the average image quality scores of 1.7 (study Site 1) and 2.3 (study Site 2), respectively (1= best, 3=worst quality score).

We then investigated image quality as a confounding factor for segmentation performance (Fig. 2A) showing a significant effect for both manual segmentations (Kruskal-Wallis P1, P2, P3, p-values=[1.24×10^{-6}, 1.14×10^{-5}, 4.21×10^{-8}] , n=107) and the models accordingly (Kruskal-Wallis MP1, MP2, MP3, MV, p-values=[1.53×10^{-7}, 7.14×10^{-7}, 4.67×10^{-8}, 2.53×10^{-7}], n=107), indicating lower MRI quality to result in lower segmentation performance.

Model robustness was furthermore tested for lung structural differences, i.e., presence of BPD-characteristic changes (Fig 2B), showing no significant differences for segmentation performance between disease severity grades (Kruskal-Wallis MP1, MP2, MP3, MV, p-values=[0.30, 0.20, 0.55, 0.48], n=107).

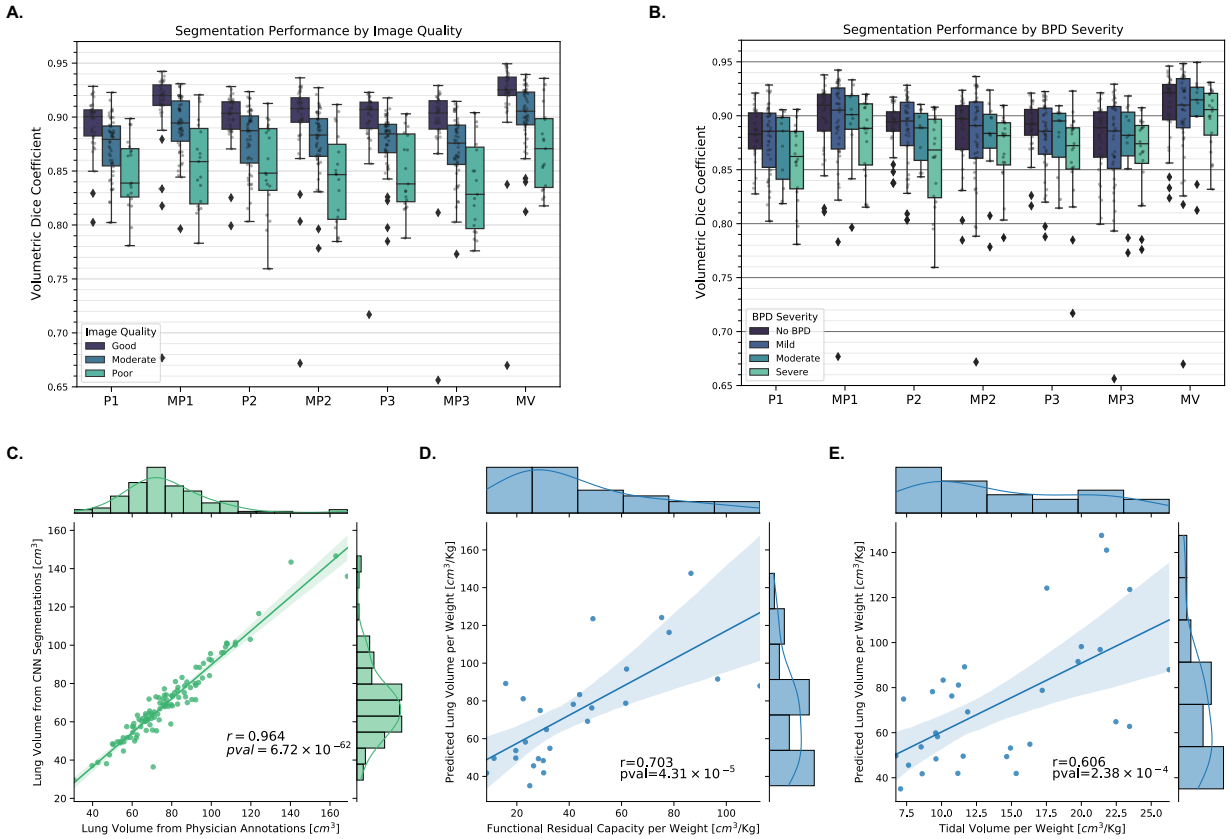


Figure 2. Lung Segmentation and Lung Volume Analysis. (**A**) Image quality vs segmentation performance for manual annotations and models (n=107). (**B**) Manual annotations and model segmentation performance by BPD Severity (mild, moderate,

and severe) (n=107). (**C**) Estimated lung volume from the CNN lung segmentations vs estimated lung volume from manual segmentations (n=107). (**D**) Correlation of functional residual capacity per weight vs predicted lung volume per weight (n=27). E) Correlation of tidal volume per weight vs lung volume per weight (n=32).

2.3 MRI-based Lung Volume Estimates Correlate Well with Lung Function

In order to approximate lung volume, a 3D reconstruction of the lungs was performed using the fully segmented lung sequences (*Methods - MRI-Lung Volumetric Features*). We classified the lung volume further by addressing each lung lobe separately using an automatic clustering technique allowing for the differentiation of the lung lobe volume ratio.

Predicted lung volumes showed a significant correlation (Pearson, r=0.964, p-value=6.72×10^{-62}, n=107) when compared to the lung volumes generated by manual segmentations (Fig. 2C), indicating that the high segmentation accuracy of the DL ensemble model enabled a robust downstream estimation of the lung volumes, including sequences with low image quality (Pearson, r=0.963, p-value=3.69×10^{-11}, n=19).

A significant correlation was also observed between the estimated lung volume normalized to bodyweight and parameters derived from infant lung function testing (ILFT by bodyplethysmography), such as functional residual capacity normalized to bodyweight (Fig. 2D, r=0.703, p-value=4.31×10^{-5}, n=27), and tidal volume normalized to bodyweight (Fig. 2E, r=0.606, p-value=2.38×10^{-4}, n=32), thereby validating the accuracy of the automated volume prediction by an MRI-independent measure.

2.4 MRI-based Lung Features Demonstrate Predictive Performance for BPD Severity Classification

To investigate the clinical value of the MRI-based lung volume in the preterm neonate, we tested the performance of ML models for BPD severity prediction.

First, an exploratory analysis of the relation between the predicted lung volumes and indicators of BPD severity was performed. We demonstrated significant differences between the lung volume per bodyweight distributions for different BPD severity grades (Kruskal-Wallis, k=43.86, p-value=1.61×10^{-9}, n=103), with higher lung volumes corresponding to increased BPD severity levels (Wilcoxon–Mann–Whitney U test with Bonferroni correction, no-BPD vs mild, no-BPD vs

moderate, no-BPD vs severe, mild vs moderate and mild vs severe, k=[226, 25, 57, 98, 210], p-values=[1.22×10^{-5}, 1.18×10^{-4}, 2.86×10^{-6}, 3.30×10^{-2}, 2.03×10^{-2}], n=103; Fig. 3A). Moreover, significant positive correlations were observed between the predicted lung volume normalized to bodyweight and the duration of mechanical ventilation (invasive and non-invasive) (r=0.738, p-value=5.54×10^{-19}, n=103; Fig. 3B), as well as, the duration of oxygen supplementation (r=0.622, p-value=2.39×10^{-12}, n=103; Fig. 3C). These findings thereby confirm the positive correlation demonstrated for formerly preterm infants at school age showing an increase in functional residual capacity in higher BPD severity grades and a prolonged history of respiratory support [14], in line with previous observations reporting elevated lung volumes in severe BPD cases [7,10].

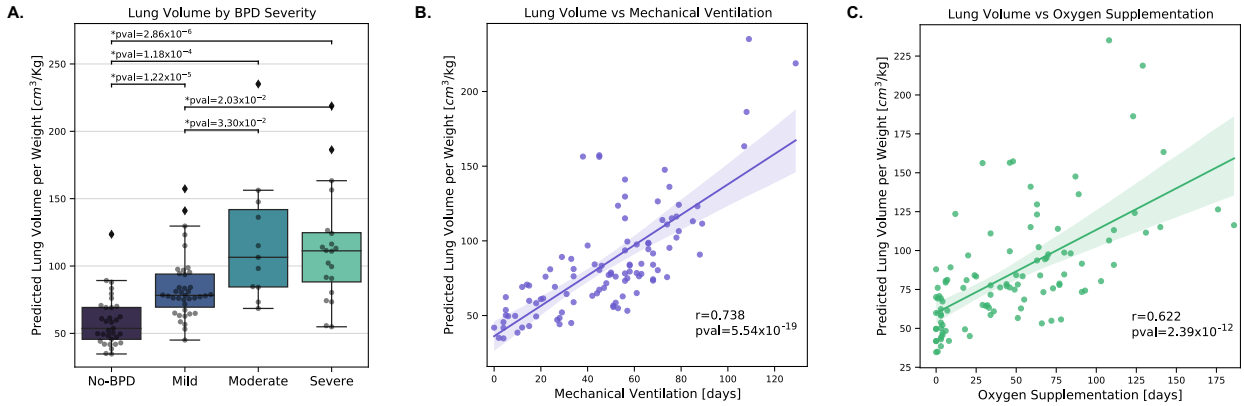


Figure 3. Lung Volume vs BPD Severity Indicators. (**A**) Distribution of the predicted lung volume normalized by bodyweight against BPD severity grades. (n=103, **p-values* for Wilcoxon–Mann–Whitney U-test with Bonferroni correction). (**B**) Correlation of lung volume normalized by bodyweight vs duration of mechanical ventilation (n=103). (**C**) Correlation of lung volume normalized by bodyweight vs duration of oxygen supplementation (n=103).

Next, MRI-based lung volumes normalized to bodyweight as well as lung-lobe ratios were used as explanatory features to predict BPD severity and days of mechanical ventilation using Logistic Regression (LR), Random Forest (RF), and Poisson regression. Additional clinical variables were sequentially added as features to investigate the overall predictive performance of the model. Ten repeated nested cross-validations were performed with five outer and five inner-folds for performance estimation and hyperparameter tuning. The dataset used for the predictions consisted of 103 patients with a complete set of explanatory variables. Three prediction scenarios were explored with our models: i) a binary classification model comparing infants with BPD and without BPD diagnosis including all severity levels, ii) a scenario with multinomial classification

comparing the three different BPD severity levels to infants without BPD, and iii) a regression scenario predicting the number of days with mechanical ventilation as a continuous measure indicating BPD severity.

For the binary classification, a high classification performance was found with lung-volumetric features (V) (Table 1) indicating an average AUC performance of 84.31%±8.66% for the LR Model and 82.66%±10.64% for the RF Model. The inclusion of patient information (P) increased the AUC classification performance to 89.40%±7.95% for the LR model and 89.17%±7.70% for the RF model (Fig. 4A-B). Further inclusion of clinical variables (C) showed similar AUC performances of 89.51%±8.20% for the LR model and 88.89%±7.95% for the RF model (Fig. 4A-B). F1 scores were higher for the RF models with 85.10% ±6.11% for the highest F1 score using all feature groups (Table 1). The ROC curves for the binary classification LR model with all feature sets (V,P,C) also showed an overall stable model for the different train-test splits (Fig. 4C). The high performance of the classification models that exclusively used volumetric features indicates their potential as descriptors of BPD severity, which can be improved to an overall high performance when adding known clinical risk factors, such as gestational age (GA) and growth [15,16].

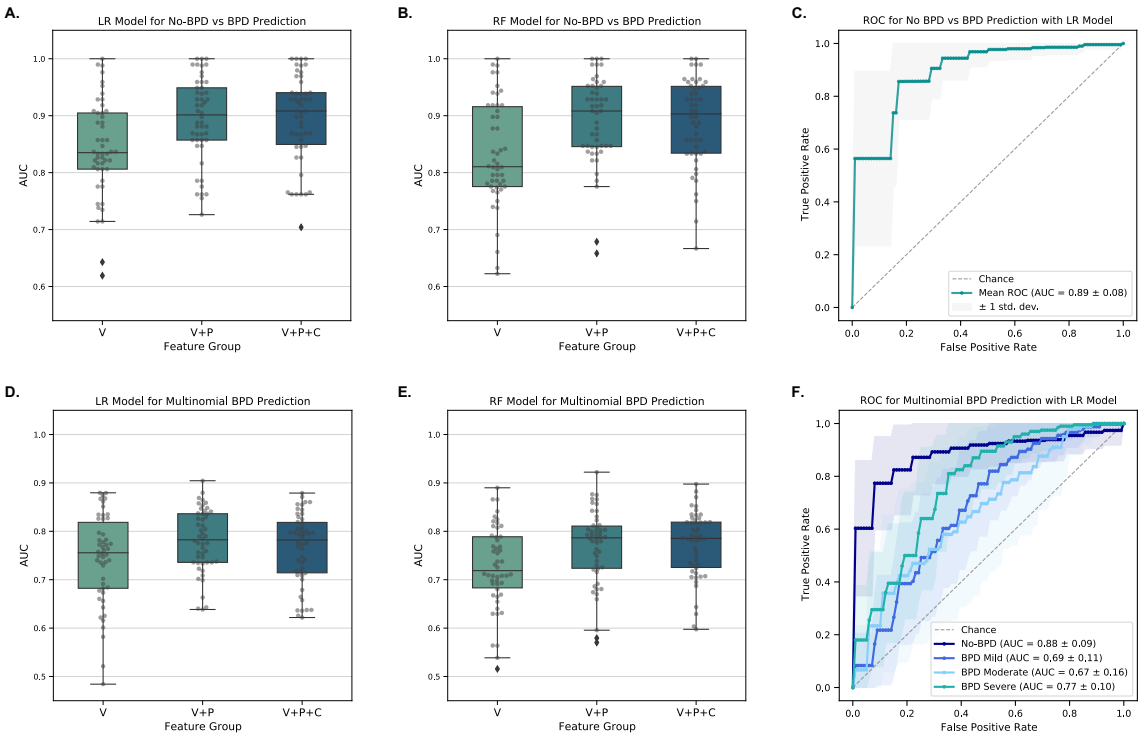


Figure 4. BPD Prediction Performance. V= Volumetric features (lung volume per bodyweight, lung lobe volume ratio), P= Patient information (gender, gestational age, birth weight, body Size), C= Clinical Parameters (APGAR 5 min score, early onset infection, steroids treatment). (**A**) AUC performance of logistic regression models for BPD prediction (BPD vs. no BPD) by feature groups. (**B**) AUC performance of random forest models for BPD prediction (BPD vs. no BPD) by feature groups. (**C**) ROC for BPD binomial classification with logistic regression and all feature groups (V+P+C). (**D**) AUC performance of logistic regression models for multinomial BPD prediction by feature groups. (**E**) AUC performance of random forest models for multinomial BPD prediction by feature groups. (**F**) ROC for BPD multinomial classification with logistic regression using all feature groups (V+P+C).

The multiclass prediction of BPD that only used volumetric features as explanatory variables resulted in an pairwise average AUC of 74.33%±9.30% with the LR model, and a pairwise average AUC of 72.65%±8.49% for the RF model (Fig. 4 D-E, Fig. S1). The best performing model for multinomial classification was the LR model using features from volumetric measurements and patient-information with a pairwise average AUC of 78.06%±6.51% (Table 2). The ROC of the multinomial LR classification using all features showed that the best one class vs all AUC performances can be reached for no-BPD and severe BPD classes (Fig. 4F).

Table 2 - Binary and Multinomial BPD Severity Classification Performance

		Logistic Regression Model			**Random Forest Model**		
	Feature Group → ↓ Score	**V**	**V + P**	**V+P+C**	**V**	**V + P**	**V+P+C**
Binary: No BPD vs BPD (all severity levels)	AUC [%]	84.31 ±8.66	89.40 ±7.95	89.51 ±8.20	82.66 ±10.64	89.17 ±7.70	88.89 ±7.95
	Weighted F1 Score [%]	81.54 ±8.29	83.71 ±7.35	83.32 ±8.20	80.22 ±8.25	84.71 ±5.78	85.10 ±6.11
Multinomial: No BPD, BPD Mild, Moderate and Severe.	AUC [%]	74.33 ±9.30	78.06 ±6.51	76.44 ±7.35	72.65 ±8.49	76.84 ±7.69	77.22 ±6.92
	Weighted F1 Score [%]	46.75 ±8.60	50.71 ±7.22	52.31 ±8.13	48.97 ±10.37	52.94 ±9.49	53.87 ±8.67

V= Volumetric features (lung volume per bodyweight, lung lobe volume ratio), **P=** Patient information (gender, gestational age, birth weight, body Size), **C=** Clinical Parameters (APGAR 5 min score, early onset infection, steroids treatment).

The prediction of days of mechanical respiratory support was evaluated with RF regression, and Poisson regression, next to an analysis by feature groups (Table 3). The RF regression model achieved the lowest mean average prediction error (MAE) with 14.15 ±2.13 days using only

volumetric features, 10.87 ±1.61 days with the (V+P) feature groups, and 10.78 ±1.76 days for the (V+P+C) feature groups.

Table 3 - Mean Average Prediction Error of Days with Mechanical Respiratory Support

	Mean Average Error (MAE) in days by Feature Group		
	V	**V + P**	**V+P+C**
Random Forest Regression	14.15 ±2.13	10.87 ±1.61	10.78 ±1.76
Poisson Regression	17.40 ±2.93	13.11 ±2.75	12.67 ±1.99

V= Volumetric features (lung volume per bodyweight, lung lobe volume ratio), **P=** Patient information (gender, gestational age, birth weight, body Size), **C=** Clinical Parameters (APGAR 5 min score, early onset infection, steroids treatment).

3. Discussion

In order to improve standardized image assessment and thus diagnostic accuracy for lung disease in high-risk patient cohorts, we successfully applied DL models to demonstrate the viability for accurate segmentation of neonatal lung in MRI sequences thereby addressing the most challenging conditions. The high comparability and low variability of the CNN models in comparison to the manual annotations, implies the significant potential of the models to overcome the technical challenges of newborn quiet-breathing MRI, including small volumes, motion artifacts, blurring, and low image resolutions. Previous studies that performed MRI lung segmentation in neonates faced limitations in scalability and sensitivity due to smaller cohorts and the use of shape-based image-segmentation methods. For instance, Heimann et al. [17] used lung shape-appearance models to perform free-breathing MRI lung segmentation in a cohort of 32 children reporting an average volumetric overlap of 85% with the annotated ground truth. In adult subjects, MRI acquisition protocols are improved to meet the need of such automated approaches by the use of breath holding maneuvers, impossible in the spontaneously breathing infant. By the use of this technique, Kohlmann et al. achieved a ground truth segmentation overlap of 94% using thresholding and 3D lung region-growing-based methods with 14 patients [18], whereas other adult MRI lung segmentation methods reported VDCs in the range of (82%-86%) [19, 20]. In comparison, our ensemble DL model applied in the most challenging of conditions achieved a significant performance with an average VDC of 90.2%, while using a multi-center approach.

Furthermore, the equivalent segmentation performance in healthy and diseased lungs indicates persistent accuracy in different lung structure-related image conditions for the proposed 2D segmentation method even when structural differences apply.

As a result of the high segmentation performance of the proposed ensemble model, our automated pipeline also enabled an accurate downstream estimation of the neonatal lung volumes, significantly correlated to the corresponding volumes abstracted from manual annotations (r=0.964). Despite the significant technical challenges faced in neonatal lung MRI, our results thereby reach comparable performance levels reported for adult cohorts, where MRI lung volumetric estimations were reported with Pearson correlations above 0.98 when comparing manual vs automatic volume predictions [18]. Furthermore, our MRI-based automatic volume estimations demonstrate high consistency when compared with lung function measurements, providing a valuable validation independent from image-based annotations.

Next, we show that the predicted lung volume and lung-lobe volume ratio hold potential to reflect lung health and disease, i.e., BPD. Here, the significant performance of the BPD classification and regression models that include volumetric features indicates their significant value for disease prediction (AUC 0.895±0.082), thereby exceeding previous imaging-based BPD prediction models that reported AUC binary prediction performances of 0.834-0.858 using ultrasound, and an AUC of 0.8 when using MRI time-relaxation periods [21]. Hence, our results motivate further research in the application of automated segmentation and extraction of lung-volumetric features for monitoring infant lung diseases, potentially while integrating additional lung structural information [21].

Further improvements could be achieved by future studies through the collection of larger and more diverse datasets of manual annotations to strengthen the generalizability and performance of the ensemble model, next to the inclusion of different lung pathological conditions. Investigation of additional volumetric and spatial features that relate to the characterization of the pulmonary condition of the lung, together with the analysis of MRI-lung volumetric features in longitudinal approaches could become crucial to inform medical decision making through early prediction of outcome.

Our work herewith contributes to the generation of AI-driven scientific evidence required to integrate MRI volume-based features as a biomarker to monitor neonatal lung development in health and disease in daily clinical practice while avoiding radiation exposure. The proposed DL segmentation method and automated extraction of structural measurements from neonatal lung MRI, enables the translation of medical expertise to larger scale applications, including its transferability to health centers that face different expertise levels as well as longitudinal measurements over prolonged periods of time. Therefore significantly contributing to the standardization and comparability of critical features in respiratory disease monitoring in newborns and infants.

4. Materials and Methods

4.1 Cohort Characteristics

A total of 107 preterm infants, gestational age (GA) 27 ±2.13 weeks, with and without BPD, were prospectively included in the study from two medical centers after informed parental consent: the Perinatal Centre LMU Munich (Site 1, n=86; EC #195–07) and the Perinatal Centre UKGM Giessen (Site 2, n=21; EC #135–12). In total, 73 of the participants were diagnosed with BPD and classified into three severity levels: mild (n=42), moderate (n=11) and severe (n=20), according to the definition by Jobe et. al. [1], 34 participants did not develop BPD. Clinical information on neonatal health conditions and treatments was also collected from both cohorts, all the clinical variables were available for 103 patients of the complete cohort (Table 1). Pulmonary function tests including tidal breathing analysis and bodyplethsymographic functional residual capacity, were performed for a subgroup of neonates (n=32) at 36 weeks GA, according to the guidelines of the American Thoracic and European Respiratory Society.

Table 1 - Clinical Information of the Preterm Neonatal Cohort (N=103)

Clinical Variable	**All (N=103)** Average ±SD	**No-BPD (N=33)** Average ±SD	**BPD-Mild (N=39)** Average ±SD	**BPD-Moderate (N=11)** Average ±SD	**BPD-Severe (N=20)** Average ±SD
Gestational Age (weeks)	26.96 ±2.12	29.09 ±1.43	26.20 ±1.48	25.69 ±2.06	25.62 ±1.43

Birth Weight (g)	908.25 ± 304.58	1206.21 ±292.39	829.74 ± 182.77	641.82 ± 177.85	716.25 ± 154.31
*Body Size (cm)	34.38 ± 4.02	38.38 ±3.23	33.26 ± 2.92	31.06 ±2.43	31.76 ±2.25
APGAR Score - 5 min	7.71 ± 1.40	8.06 ±1.00	7.87 ±1.10	7.36 ±2.38	7.00 ±1.59
†Early Onset Infection	No (N=80), Yes (N=23)	No (N=29), Yes (N=4)	No (N=30), Yes (N=9)	No (N=7), Yes (N=4)	No (N=14), Yes (N=6)
Administration of post-natal corticosteroids	No (N=61), Yes (N=42)	No (N=28), Yes (N=5)	No (N=22), Yes (N=17)	No (N=6), Yes (N=5)	No (N=15), Yes (N=5)
Oxygen Supplementation (days)	47.30 ± 43.18	5.18 ±7.72	45.56 ±21.23	81.55 ±30.96	101.35 ±40.78
Mechanical Ventilation (days; invasive and non-invasive)	48.22 ±26.93	19.91 ±15.62	52.51 ±13.74	66.55 ±19.20	76.50 ±21.08

*Linear BMI imputation performed for missing body sizes. †Early Onset Infection as defined by [22].

4.2 Imaging and Segmentation Protocols

Preterm infants underwent MRI near term age, i.e., at approximately 36 weeks GA. T2-weighted lung MRI sequences were acquired in unsedated spontaneous sleep for the Perinatal Centre LMU Munich cohort, and under light sedation with chloral hydrate (30-40 mg/kg administered orally) for Perinatal Centre UKGM Giessen.

Axial images were acquired with a T2-weighted half-Fourier-acquired single-shot turbo spin echo (HASTE) protocol for lung structural assessment. An ECG-triggered 2D multi-slice single-shot fast spin-echo sequence with an echo time (TE) of 57 ms was used; the repetition time was set to 2 RR intervals. The spatial resolution was 1.3×1.9 mm² in plane with a slice thickness of 4 mm and 0.4 mm slice gap. Parallel imaging with an acceleration factor of 2 was applied and 2 averages were acquired for each slice.

In sum, a total of 107 MRI sequences with 2,165 axial images, with a resolution of 256×192 pixels, were acquired using 3T MRI scanners (Siemens Skyra for the Perinatal Centre LMU Munich and Siemens Verio for Perinatal Centre UKGM Giessen). Pseudonymization of image and clinical information was performed to guarantee a blinded analysis.

Manual lung segmentation of the MRI sequences was performed independently by three physicians, with different training levels (one radiologist and two late-stage image analysis trained

medical students). The software ITK-SNAP [23] was used to collect the manual segmentations. Image quality of the sequences was rated by a fourth independent radiologist. To remove unnecessary background, we first identify the centroid of all pixels that are above the 5% intensity quantile threshold across all slices and then crop all slices to a square of 128×128 pixels centered at the centroid.

4.3 Deep-Learning MRI Lung Segmentation Model

The proposed 2D lung segmentation DL models are based on the U-Net neural network (NN) architecture [24]. U-Net models produce a latent representation of the image by processing it through a set of convolutional layers in a contracting path and then processing the features through an expansion path of up-convolutional layers, with skipped connections at each level, returning a high-resolution binary pixel-wise segmentation map of the image. Our U-Net architecture has 4 down and 4 up convolutional blocks and a fifth intermediate convolutional block, batch normalization was included after every building block of the U-Net and a Dropout Layer (dropout rate=0.1) after each econding block was added. Detailed architecture parameters are available in Table S3 and in the code-repository (*Section 4.6*). The Instant-DL framework, which is designed to efficiently train U-Net segmentation models for medical imaging applications, was adapted for our study [25]. Hyperparameters were optimized using grid search (Table S2) for three randomly selected leave-one-patient-out models in a 4-fold cross-validation scheme, the best performances were achieved with 300 training epochs, 0.001 learning rate, using a binary cross-entropy loss for the NN optimization, and applying image augmentations with 0.1 random zoom, 0.1 translations and up to 22.5° random rotations.

The dice-coefficient, defined as $DC = \frac{2pg+1}{p^2+g^2+1}$, with p being the predicted positive-class pixels and g being the ground truth pixels [26], was used as the metric for evaluation of the model performance. Ground truths were generated with the annotations from a single physician, for the physician-based models, and with majority voting from all the physicians, for the integrated model. Optimization was performed with Adam [27].

Using a leave-one-patient-out cross-validation scheme, a set of k models were trained exclusively with the sequences from Site 1, that is, for each kth model, the data of the kth participant is used

395 only for validation. In addition, MRI sequences from study Site 2 were used exclusively for
396 validation of a MV model trained with all the sequences from study Site 1. Lung segmentation
397 accuracy was measured using the volumetric dice coefficient (VDC), where pixels from all the
398 slices in the MRI sequence are aggregated and evaluated for segmentation validity. Resulting VDC
399 scores of the models and manual annotations for each sequence can be found on Table S1.

400

401 *4.4 MRI-Lung Volumetric Features*

402 A 3D volumetric representation of the lung was created using the segmented lung regions and the
403 DICOM pixel-spacing and patient orientation metadata from the MRI 2D sequences. The volume
404 was generated with voxels that considered both the distance to the neighboring pixels in the 2D
405 slice (dx, dy), as well as the slice thickness and the space between the slices (dz). The total lung
406 volume was calculated by adding the individual voxel volumes of all the segmented pixels in the
407 MRI sequence.

408 In addition, the lung was divided between the left and right lobes using a two-step algorithm
409 involving a K-means (K=2) clustering with further refinement of the class labels using a soft-
410 margin SVM classifier (penalty=0.001). For the K-means algorithm, we initialized the centroids
411 using the *k-means++* initialization [28] and used a weighted Euclidean distance favoring the x and
412 y dimensions for improved left-right lung lobe clustering with weights of (1, 1, 0.1) for x, y, and
413 z coordinates. Using k-means annotations as labels, the SVM algorithm was applied iteratively
414 updating the voxel labels until convergence of the resulting silhouette score; methods are available
415 in the code-repository (*Section 4.6*). Once the voxels were classified in the left or right lung lobes,
416 the lung lobe volume ratio was calculated by finding the volume of each lobe and then by dividing
417 the larger over the smaller lobe volume as an asymmetry indicator. The resulting MRI-based lung
418 volume features are available in Supplementary Table S1.

419

420 *4.5 BPD Severity Prediction Models*

421 A regression analysis to predict the severity of BPD was performed for which explanatory
422 variables were grouped in three categories, MRI-based lung volume features (V) (lung volume per
423 birth weight and the lung lobe volume ratio), patient-related (P) features (gestational age, birth

weight, body size, gender) and also clinical parameters (C) (APGAR score - 5 min, early onset infection, steroid treatment).

Random Forest (RF) [29] and Logistic regression models [30] were trained to perform binomial (BPD vs. No BPD) and multinomial (no BPD, mild, moderate, and severe BPD) classification. A nested cross-validation scheme was implemented to find the best hyperparameters using grid-search (Table S2), the average performance of the model was estimated with multiple repetitions of the nested cross-validation scheme (10 times with different random seeds). A stratified 5-fold train-test split was used for both the inner and outer loops of each nested cross-validation. For the Logistic regression model, features were standardized removing the mean and scaling to unit variance of the training set. Additional regression models, Poisson and RF, to predict the number of days with required respiratory support (adding invasive and non-invasive days) were trained using the same nested cross-validation scheme.

4.6 Data and Source Code Availability

Source code of models for lung segmentation, 3D volume-feature estimations, and regression models can be found at https://github.com/SchubertLab/NeoLUNet.

Resulting weights of the U-Net models used for BPD prediction will be made available at (https://zenodo.org).

5. Acknowledgments

We sincerely thank the patients and their families of the AIRR study cohort for their significant contribution to the study by providing the samples. The present study was supported by the Young Investigator Grant NWG VH-NG-829 by the Helmholtz Foundation and the Helmholtz Zentrum Muenchen, Germany and the German Center for Lung Research (DZL) - German Ministry of Education and Health (BMBF).

B.M. and A.C. are supported by the Helmholtz Association under the joint research school Munich School for Data Science - MUDS. B.S. acknowledges financial support by the Postdoctoral Fellowship Program of the Helmholtz Zentrum München.

453 **REFERENCES**

454 [1] A. H. Jobe and E. Bancalari, "Bronchopulmonary dysplasia," Am. J. Respir. Crit. Care
455 Med., vol. 163, no. 7, pp. 1723–1729, Jun. 2001.

456 [2] A. H. Jobe, "Mechanisms of Lung Injury and Bronchopulmonary Dysplasia," Am. J.
457 Perinatol., vol. 33, no. 11, pp. 1076–1078, Sep. 2016.

458 [3] T. R. Kalikkot, M. C. Guaman, and B. Shivanna, "Bronchopulmonary dysplasia: A review
459 of pathogenesis and pathophysiology," Respir. Med., vol. 132, Nov. 2017, doi:
460 10.1016/j.rmed.2017.10.014.

461 [4] G. R. Washko, "Diagnostic imaging in COPD," Semin. Respir. Crit. Care Med., vol. 31,
462 no. 3, pp. 276–285, Jun. 2010.

463 [5] C. May, M. Prendergast, S. Salman, G. F. Rafferty, and A. Greenough, "Chest radiograph
464 thoracic areas and lung volumes in infants developing bronchopulmonary dysplasia," Pediatr.
465 Pulmonol., vol. 44, no. 1, pp. 80–85, Jan. 2009.

466 [6] B. Loi et al., "Lung Ultrasound to Monitor Extremely Preterm Infants and Predict
467 Bronchopulmonary Dysplasia. A Multicenter Longitudinal Cohort Study," Am. J. Respir. Crit.
468 Care Med., vol. 203, no. 11, pp. 1398–1409, Jun. 2021.

469 [7] L. M. Yoder et al., "Elevated lung volumes in neonates with bronchopulmonary dysplasia
470 measured via MRI," Pediatric Pulmonology. 2019, doi: 10.1002/ppul.24378.

471 [8] P. J. Critser et al., "Cardiac Magnetic Resonance Imaging Evaluation of Neonatal
472 Bronchopulmonary Dysplasia-associated Pulmonary Hypertension," Am. J. Respir. Crit. Care
473 Med., vol. 201, no. 1, pp. 73–82, Jan. 2020.

474 [9] D. C. Dean 3rd et al., "Pediatric neuroimaging using magnetic resonance imaging during
475 non-sedated sleep," Pediatr. Radiol., vol. 44, no. 1, pp. 64–72, Jan. 2014.

476 [10] L. L. Walkup et al., "Quantitative Magnetic Resonance Imaging of Bronchopulmonary
477 Dysplasia in the Neonatal Intensive Care Unit Environment," Am. J. Respir. Crit. Care Med., vol.
478 192, no. 10, pp. 1215–1222, Nov. 2015.

479 [11] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical
480 Image Segmentation: Achievements and Challenges," J. Digit. Imaging, vol. 32, no. 4, pp. 582–
481 596, Aug. 2019.

482 [12] J. Islam and Y. Zhang, "Towards Robust Lung Segmentation in Chest Radiographs with
483 Deep Learning," arXiv [cs.CV], Nov. 30, 2018.

484 [13] B. Ait Skourt, A. El Hassani, and A. Majda, "Lung CT Image Segmentation Using Deep
485 Neural Networks," Procedia Comput. Sci., vol. 127, pp. 109–113, Jan. 2018.

486 [14] J. S. Landry, T. Chan, L. Lands, and D. Menzies, "Long-term impact of bronchopulmonary
487 dysplasia on pulmonary function," Can. Respir. J., vol. 18, no. 5, pp. 265–270, Sep. 2011.

488 [15] A. P. Popova, "Mechanisms of bronchopulmonary dysplasia," J. Cell Commun. Signal.,
489 vol. 7, no. 2, p. 119, Jun. 2013, Accessed: May 10, 2021. [Online].

490 [16] B. Thébaud et al., "Bronchopulmonary dysplasia," Nature reviews. Disease primers, vol.
491 5, no. 1, Nov. 2019, doi: 10.1038/s41572-019-0127-7.

492 [17] T. Heimann, M. Eichinger, G. Bauman, A. Bischoff, M. Puderbach, and H.-P. Meinzer,
493 "Automated scoring of regional lung perfusion in children from contrast enhanced 3D MRI,"
494 Medical Imaging 2012: Computer-Aided Diagnosis. 2012, doi: 10.1117/12.911946.

495 [18] P. Kohlmann et al., "Automatic lung segmentation method for MRI-based lung perfusion
496 studies of patients with chronic obstructive pulmonary disease," Int. J. Comput. Assist. Radiol.
497 Surg., vol. 10, no. 4, pp. 403–417, Apr. 2015.

498 [19] W. F. Sensakovic, S. G. Armato 3rd, A. Starkey, and P. Caligiuri, "Automated lung
499 segmentation of diseased and artifact-corrupted magnetic resonance sections," Med. Phys., vol.
500 33, no. 9, pp. 3085–3093, Sep. 2006.

501 [20] T. Böttger et al., "Implementation and evaluation of a new workflow for registration and
502 segmentation of pulmonary MRI data for regional lung perfusion assessment," Phys. Med. Biol.,
503 vol. 52, no. 5, pp. 1261–1275, Mar. 2007.

504 [21] K. Förster et al., "Altered relaxation times in MRI indicate bronchopulmonary dysplasia,"
505 Thorax, vol. 75, no. 2, pp. 184–187, Feb. 2020.

[22] M. P. Sherman, B. W. Goetzman, C. E. Ahlfors, and R. P. Wennberg, “Tracheal Aspirationand Its Clinical Correlates in the Diagnosis of Congenital Pneumonia,” Pediatrics, vol. 65, no. 2. pp. 258–263, 1980. doi: 10.1542/peds.65.2.258.

[23] P. A. Yushkevich et al., “User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability,” Neuroimage, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.

[24] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Lecture Notes in Computer Science. pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28.

[25] D. Waibel, S. S. Boushehri, and C. Marr, “InstantDL - An easy-to-use deep learning pipeline for image segmentation and classification.” doi: 10.1101/2020.06.22.164103.

[26] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation,” 2016 Fourth International Conference on 3D Vision (3DV). 2016, doi: 10.1109/3dv.2016.79.

[27] Diederik P. Kingma and Jimmy Ba , “ADAM: A method for stochastic optimization”. 3rd International Conference for Learning Representations, 2015.

[28] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, no. 85, pp. 2825–2830, 2011.

[29] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Machine Learning, vol. 63, no. 1. pp. 3–42, 2006. doi: 10.1007/s10994-006-6226-1.

[30] C. M. Bishop, Pattern Recognition and Machine Learning. Springer Verlag, 2006.

Supplementary Material

Figure S1: Receiver operator characteristics for Random Forest models.

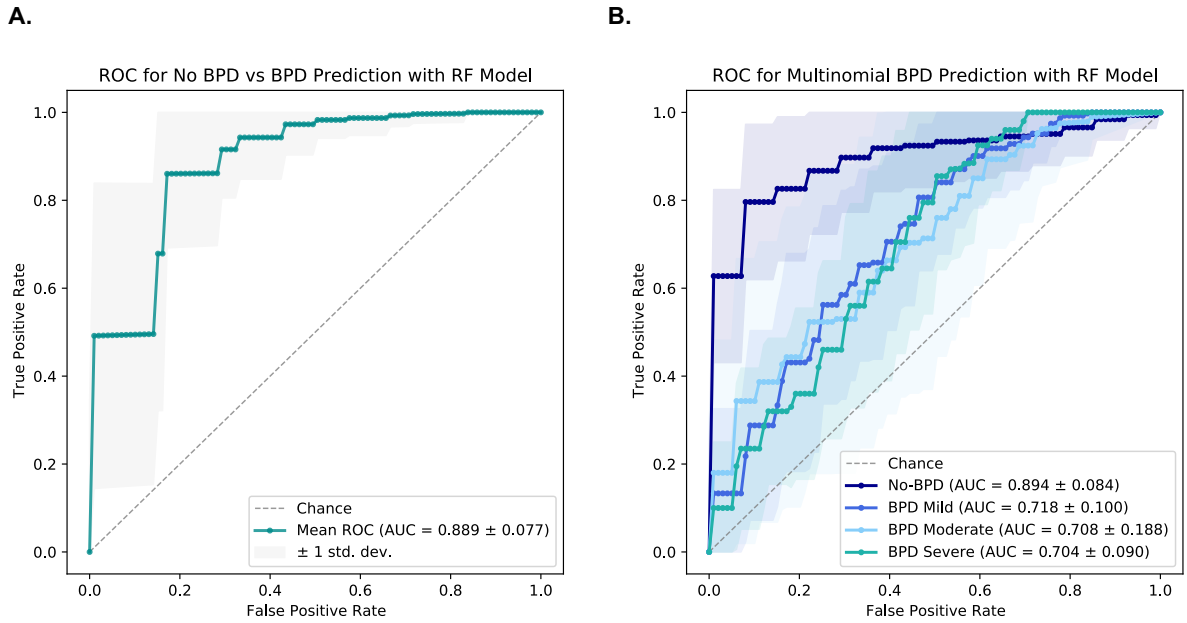


Figure S1. ROC for Binary and Multinomial RF Classification Models. (**A**) ROC for BPD binomial classification with random forest models and all feature groups (V+P+C). (**B**) ROC for BPD multinomial classification with random forest models using all feature groups (V+P+C).

Table S1: De-identified resulting physician and model VDC segmentation performances, volumetric predictions and lung-lobe volume ratio.

Table S2: Grid Search Parameters.

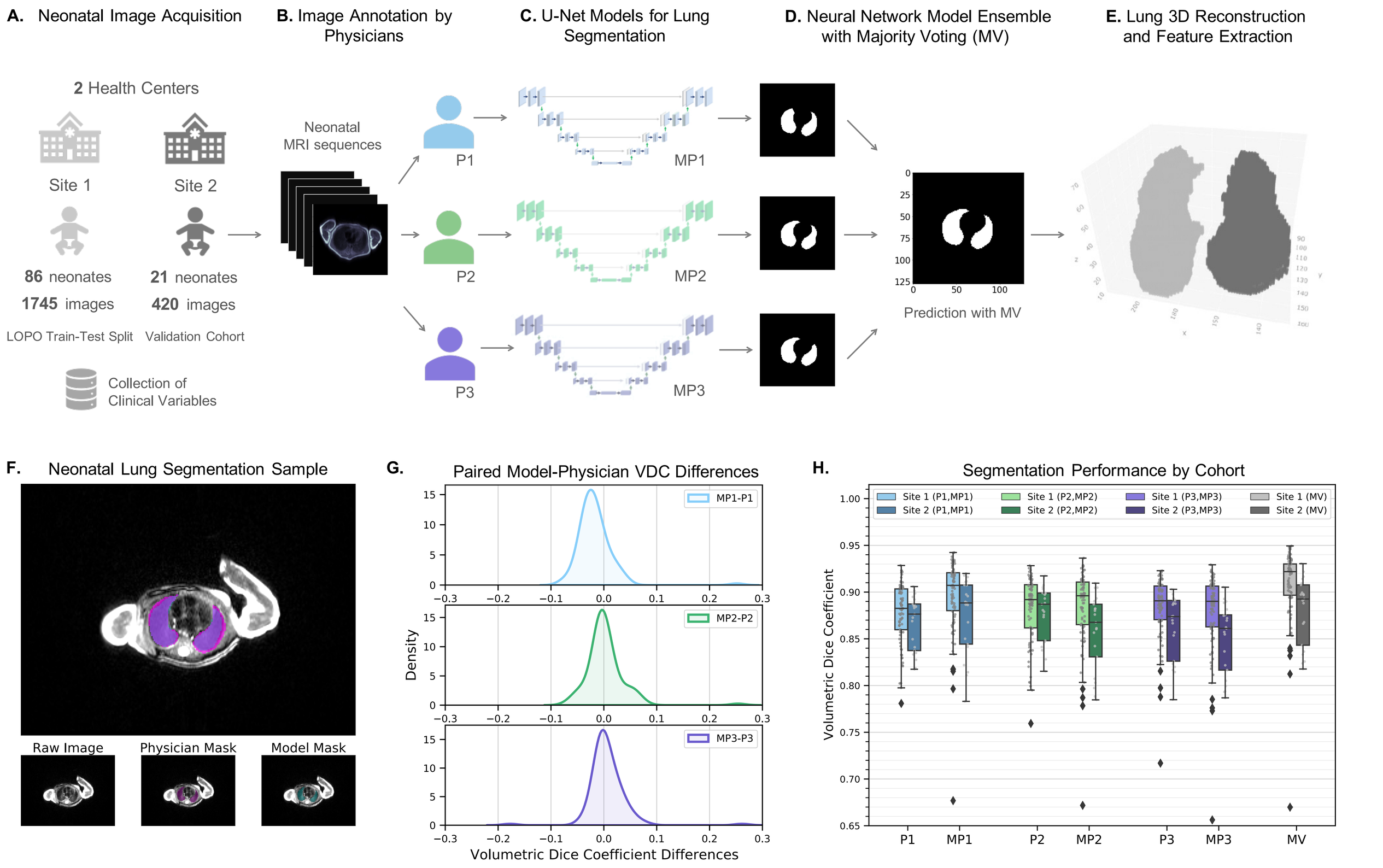
	Hyperparameters Evaluated with Grid Search
U-Net for Lung Segmentation	Epochs = [100, 200, 300, 400] Loss Functions = Mean squared error, Binary cross-entropy, Dice-loss. Learning Rate = [0.001, 0.0001] Augmentation = with augmentations (0.1 random zoom, 0.1 translations and up to 22.5° random rotations) or without augmentations.

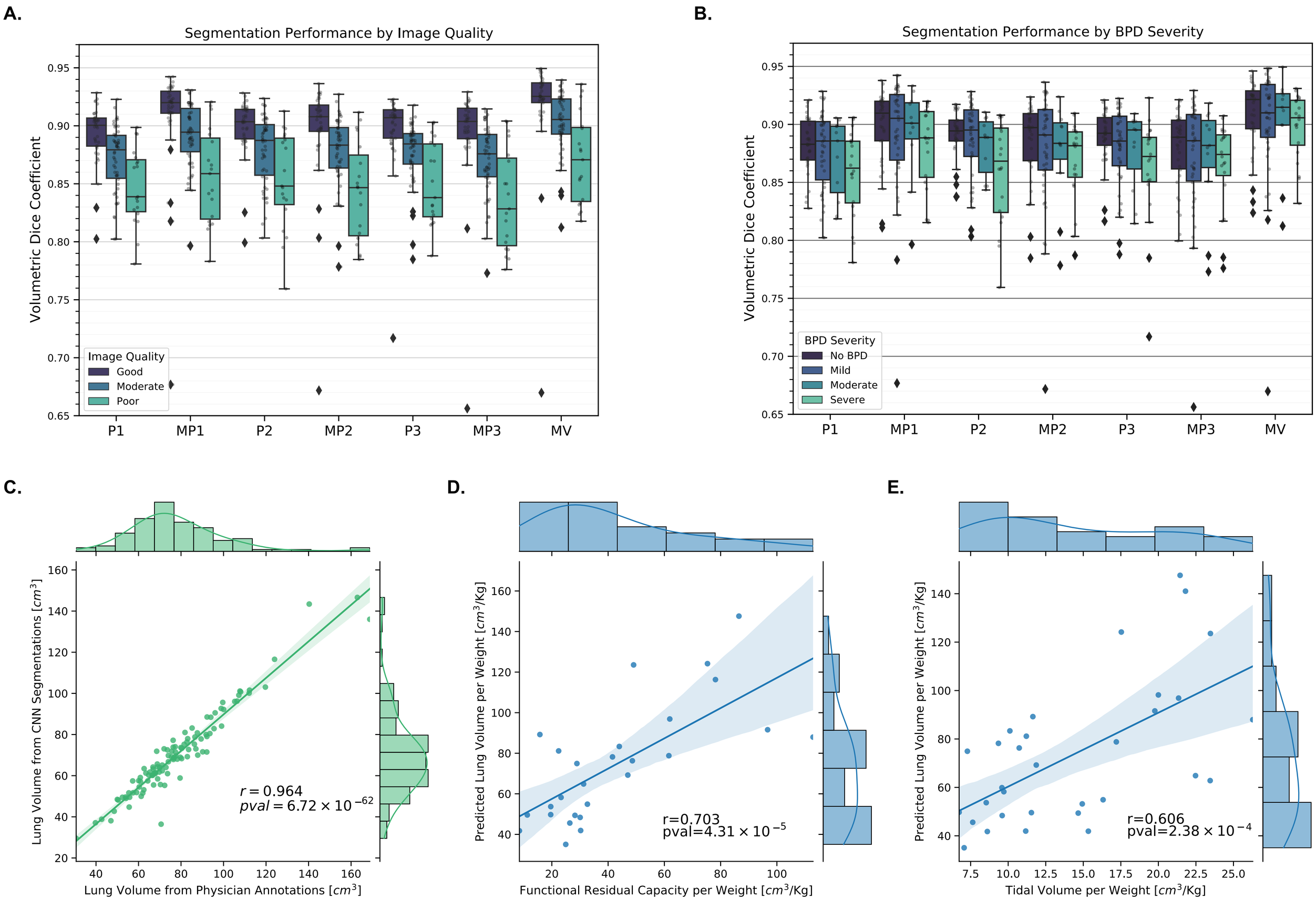
Logistic Regression	Penalty= L1, L2 C = log space sampled array (n=10 points). *np.logspace(-4, 1, 10, endpoint=True)* Grid search scoring = F1
Random Forest	Max. Depth = [3,4,5,6,7,8,9,10] Grid search scoring = F1

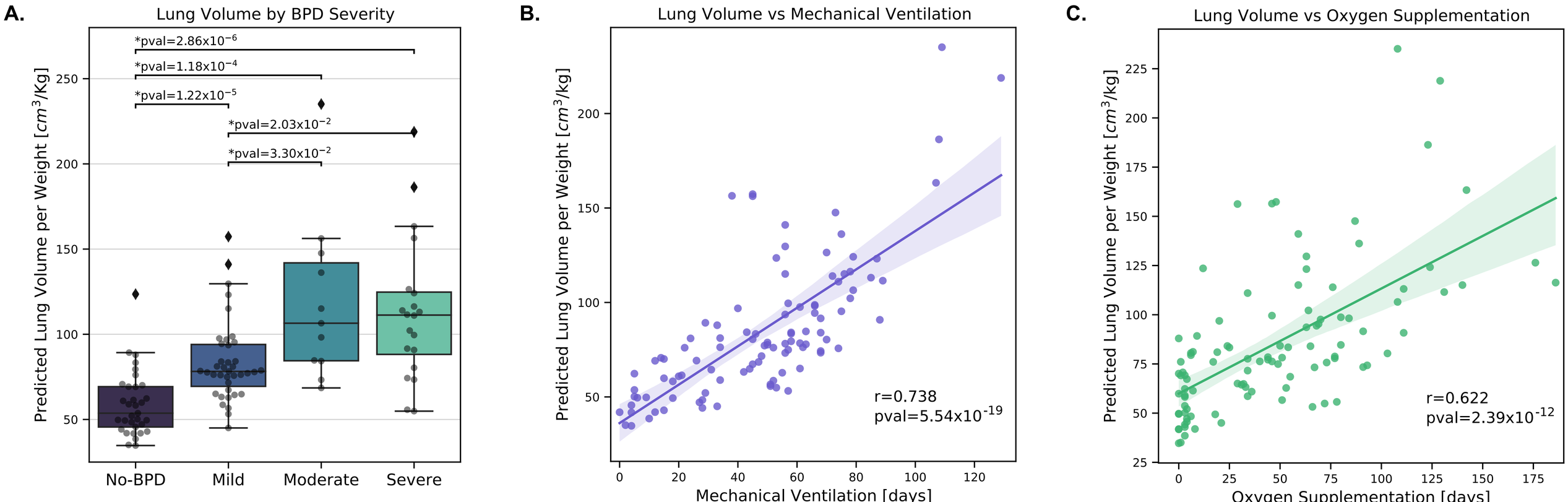
Table S3: U-Net Architecture Parameters

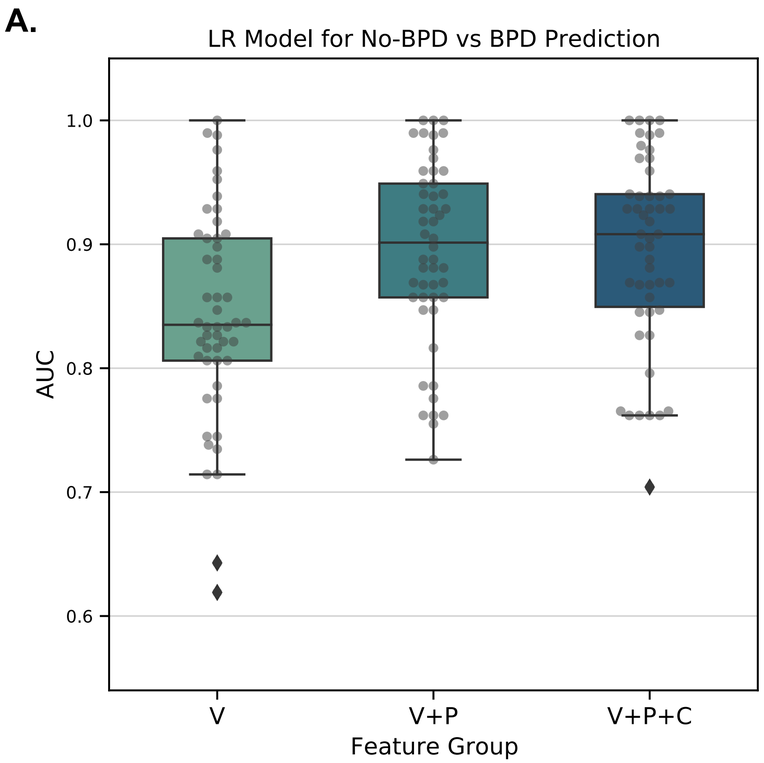
	Block Description
Convolutional Block 1 (CNN-1)	CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 2 (CNN-2)	CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 3 (CNN-3)	CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 4 (CNN-4)	CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2

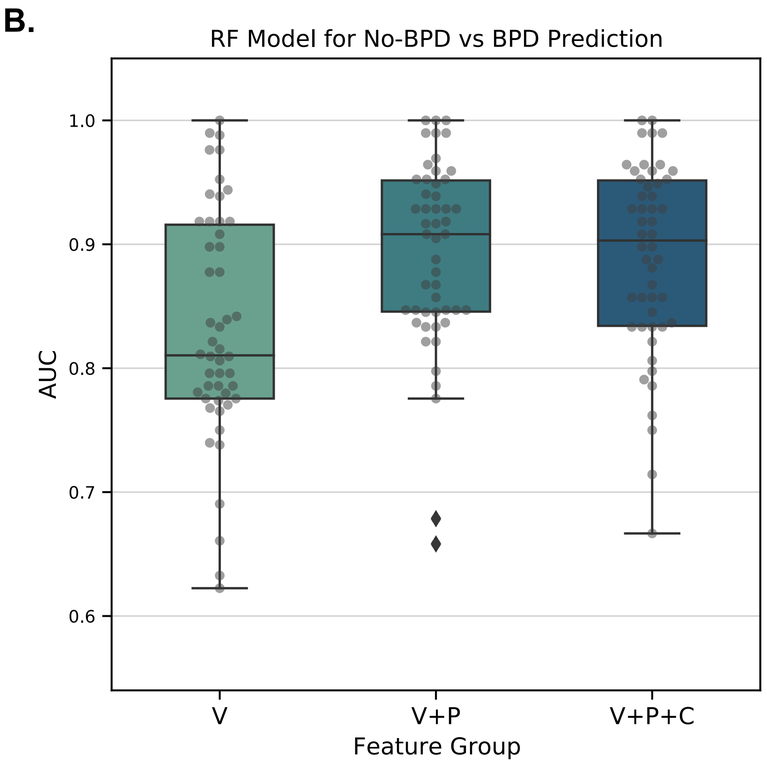
Convolutional Block 5 (CNN-5)	CNN Filters = 1024, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization CNN Filters = 1024, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 6 (CNN-6)	Up-6 Features: feature size= 512, up-sampling-kernel = 2×2 Concatenation: CNN-4 Features + Up-6 Features CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization
Convolutional Block 7 (CNN-7)	Up-7 Features: feature size= 256, up-sampling-kernel = 2×2 Concatenation: CNN-3 Features + Up-7 Features CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization
Convolutional Block 8 (CNN-8)	Up-8 Features: feature size= 128, up-sampling-kernel = 2×2 Concatenation: CNN-2 Features + Up-8 Features CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization
Convolutional Block 9 (CNN-9)	Up-9 Features: feature size= 64, up-sampling-kernel = 2×2 Concatenation: CNN-1 Features + Up-9 Features CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization
Output Layer	Activation function = Sigmoid

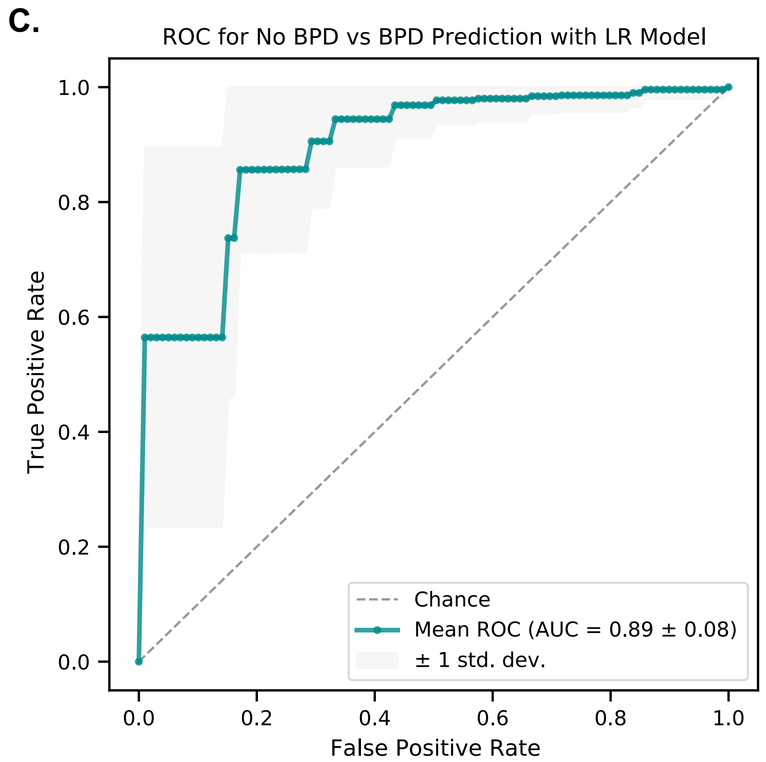

A. Neonatal Image Acquisition
B. Image Annotation by Physicians
C. U-Net Models for Lung Segmentation
D. Neural Network Model Ensemble with Majority Voting (MV)
E. Lung 3D Reconstruction and Feature Extraction
2 Health Centers
Site 1
Site 2
86 neonates
1745 images
LOPO Train-Test Split
21 neonates
420 images
Validation Cohort
Collection of Clinical Variables
Neonatal MRI sequences
P1
P2
P3
MP1
MP2
MP3
Prediction with MV
F. Neonatal Lung Segmentation Sample
Raw Image
Physician Mask
Model Mask
G. Paired Model-Physician VDC Differences
MP1-P1
MP2-P2
MP3-P3
Density
Volumetric Dice Coefficient Differences
H. Segmentation Performance by Cohort
Site 1 (P1,MP1)
Site 2 (P1,MP1)
Site 1 (P2,MP2)
Site 2 (P2,MP2)
Site 1 (P3,MP3)
Site 2 (P3,MP3)
Site 1 (MV)
Site 2 (MV)
Volumetric Dice Coefficient
P1
MP1
P2
MP2
P3
MP3
MV

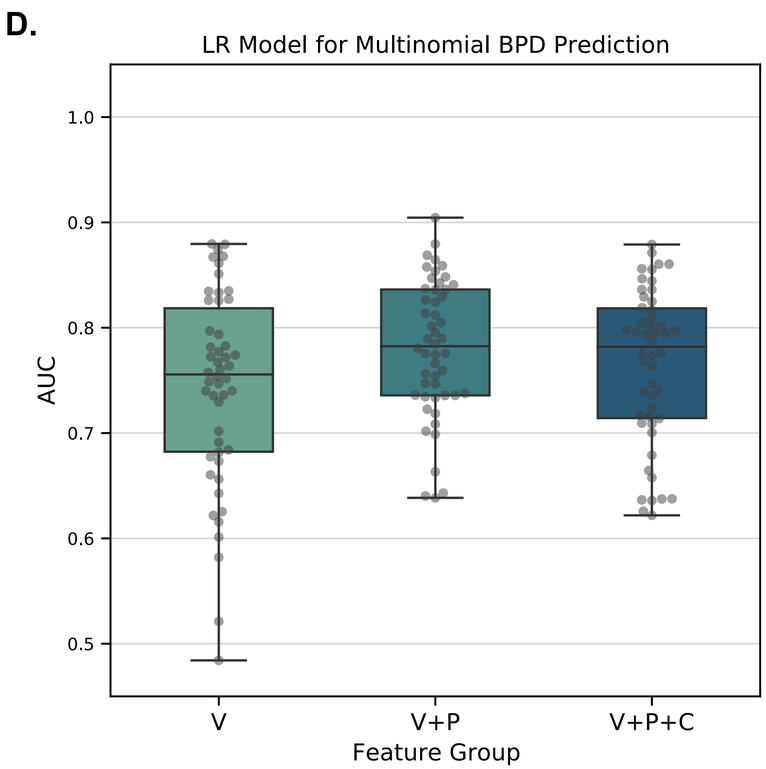

A.
Segmentation Performance by Image Quality
Volumetric Dice Coefficient
Image Quality
Good
Moderate
Poor
P1
MP1
P2
MP2
P3
MP3
MV
B.
Segmentation Performance by BPD Severity
Volumetric Dice Coefficient
BPD Severity
No BPD
Mild
Moderate
Severe
P1
MP1
P2
MP2
P3
MP3
MV
C.
Lung Volume from CNN Segmentations [cm^3]
Lung Volume from Physician Annotations [cm^3]
$r = 0.964$
$pval = 6.72 \times 10^{-62}$
D.
Predicted Lung Volume per Weight [cm^3/Kg]
Functional Residual Capacity per Weight [cm^3/Kg]
r=0.703
pval=4.31×10^{-5}
E.
Predicted Lung Volume per Weight [cm^3/Kg]
Tidal Volume per Weight [cm^3/Kg]
r=0.606
pval=2.38×10^{-4}

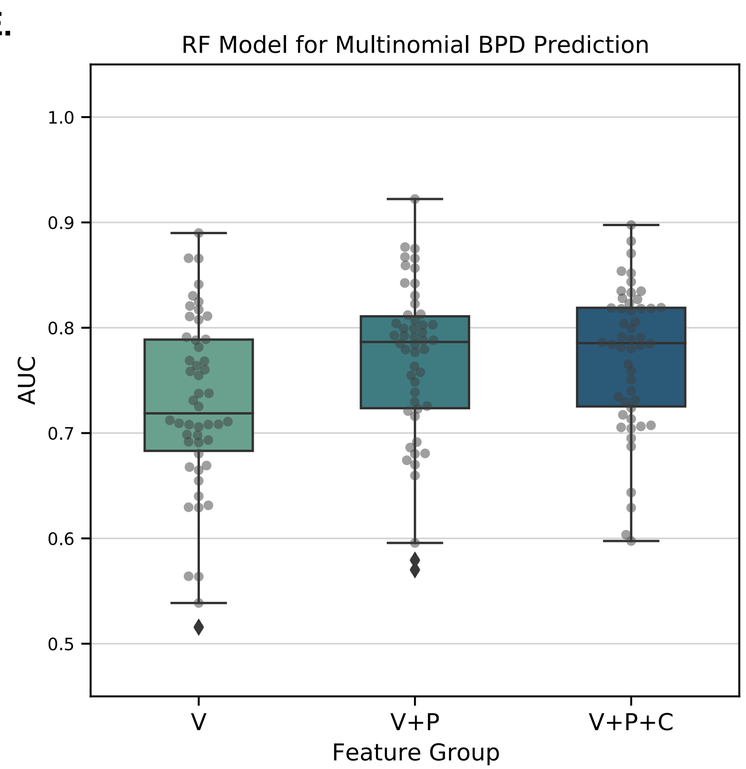

A.
Lung Volume by BPD Severity
*pval=2.86x10^{-6}
*pval=1.18x10^{-4}
*pval=1.22x10^{-5}
*pval=2.03x10^{-2}
*pval=3.30x10^{-2}
Predicted Lung Volume per Weight [cm^3/Kg]
No-BPD
Mild
Moderate
Severe
B.
Lung Volume vs Mechanical Ventilation
Predicted Lung Volume per Weight [cm^3/kg]
r=0.738
pval=5.54x10^{-19}
Mechanical Ventilation [days]
C.
Lung Volume vs Oxygen Supplementation
Predicted Lung Volume per Weight [cm^3/kg]
r=0.622
pval=2.39x10^{-12}
Oxygen Supplementation [days]

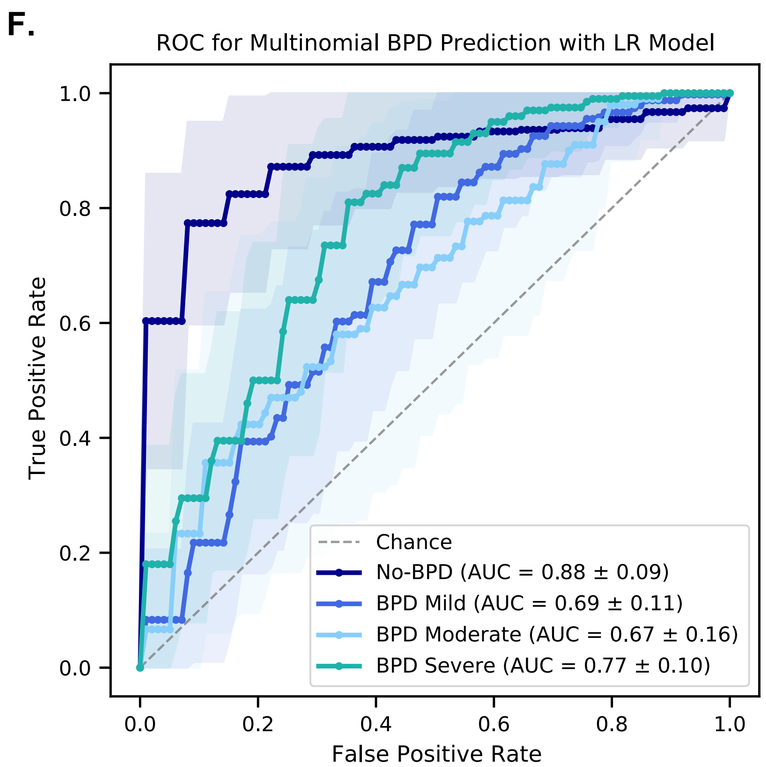

A.
LR Model for No-BPD vs BPD Prediction
AUC
1.0
0.9
0.8
0.7
0.6
V
V+P
V+P+C
Feature Group


B.
RF Model for No-BPD vs BPD Prediction
AUC
1.0
0.9
0.8
0.7
0.6
V
V+P
V+P+C
Feature Group


C.
ROC for No BPD vs BPD Prediction with LR Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Mean ROC (AUC = 0.89 ± 0.08)
± 1 std. dev.


D.
LR Model for Multinomial BPD Prediction
AUC
1.0
0.9
0.8
0.7
0.6
0.5
V
V+P
V+P+C
Feature Group


E.
RF Model for Multinomial BPD Prediction
AUC
1.0
0.9
0.8
0.7
0.6
0.5
V
V+P
V+P+C
Feature Group


F.
ROC for Multinomial BPD Prediction with LR Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Chance
No-BPD (AUC = 0.88 ± 0.09)
BPD Mild (AUC = 0.69 ± 0.11)
BPD Moderate (AUC = 0.67 ± 0.16)
BPD Severe (AUC = 0.77 ± 0.10)