

Automatic identification, segmentation, and percentage calculation of healthy and unhealthy bovine lung tissue using machine learning algorithms

Alfredo Emanuel Pavón-Pérez^{1*}, Luis Felipe Marín-Urías^{1,2}, Alejandro José Linares-Díaz^{3,4}, Carlos Morales-Morales⁴, Vaniareli Ascenez Fernández-Villa⁴, Pedro Javier García-Ramírez⁵, Manuel Barrientos-Morales⁴, Belisario Domínguez-Mancera⁴, Rogelio de Jesús Portillo-Vélez^{1,2}

¹ Faculty of Construction and Habitat Engineering, Universidad Veracruzana, Veracruz, Mexico.

² Faculty of Electrical and Electronic Engineering, Universidad Veracruzana, Veracruz, Mexico.

³ University of San Carlos de Guatemala, Chiquimula, Guatemala, Mexico.

⁴ Faculty of Veterinary Medicine and Animal Science, Universidad Veracruzana, Veracruz, Mexico.

⁵ Institute of Engineering, Universidad Veracruzana, Veracruz, Mexico.

*Corresponding author
ingfred18@gmail.com

Received: 26.04.2026

Accepted: 28.07.2026

Published: 07.09.2026

ABSTRACT. Respiratory syndromes are a major cause of morbidity and mortality in the feedlot industry, making the accurate identification of lung lesions essential for effective health management. This study proposes an automated approach for identifying and segmenting pulmonary structures and lesions using a visual machine learning model trained on a limited dataset of post-mortem images captured at a slaughterhouse. The implementation was performed using Python 3.12 with computer vision libraries and the Ultralytics framework employing YOLO11n-seg. Training incorporated both dorsal and ventral views to identify lungs, trachea and extra tissues, and lesions (designated as “findings”). Data augmentation and k-fold cross-validation were applied to address the data limitations. In addition, we developed an automated algorithm that quantifies lung health by calculating the proportion of affected versus healthy tissue through a pixel overlap comparison between segmented lung regions and detected lesions. The proposed method enables objective and reproducible assessments of pulmonary conditions from different anatomical perspectives. The results demonstrate strong performance in both anatomical segmentation and lesion detection, while also highlighting the challenges inherent in data-constrained conditions. These findings contribute to the development of computer-aided tools for veterinary diagnostics.

Keywords: artificial intelligence, automatization, classification, segmentation, visual machine learning, bovine lungs.

INTRODUCTION

Respiratory disorders in cattle are a major cause of morbidity and mortality in the beef industry, especially in intensive feedlot systems. Among these disorders, Bovine Respiratory Disease Complex (BRDC) and Acute Interstitial Pneumonia (AIP) are of primary concern because of their impact on animal health and productivity. BRDC, a multifactorial condition associated with environmental stressors, high stocking density, host immune status, and pathogen exposure (MSD Animal Health Republic of Ireland, n.d.; O'Donoghue *et al.*, 2025; Puig *et al.*, 2022), often leads to reduced performance and increased treatment costs, frequently involving the use of antimicrobials (Blakebrough-Hall *et al.*, 2020). The complex etiology of BRDC, including bacterial, viral, and/or environmental factors, complicates diagnosis and frequently results in empirical treatments, which can contribute to antimicrobial resistance. AIP is an acute and often fatal disease characterized by severe pulmonary damage and edema. Surviving animals may experience permanent impairment of lung function, reducing feed efficiency. Similar to BRDC,

AIP lacks specific clinical or pathological markers, and diagnosis relies on post-mortem and histopathological evaluations (Chien *et al.*, 2022).

Although sensor-based systems and computer vision have shown potential for the early detection of respiratory health issues in live animals (Casella *et al.*, 2023; Singh *et al.*, 2025; Wu *et al.*, 2023), post-mortem inspection at slaughterhouses remains essential for assessing lung health at the population level (Bortoluzzi *et al.*, 2023). These inspections provide valuable epidemiological data, enabling the identification of disease patterns and supporting decision-making in herd health management. This assists in the implementation of preventive measures aimed at reducing health risks and minimizing associated economic losses.

In the Mexican context, the country plays a relevant role in livestock production. The 2022 Agricultural Census conducted by the Instituto Nacional de Estadística y Geografía (INEGI) (n.d.) reported 24.5 million cattle, with 12.1% raised in fattening systems. More than half of the national production is concentrated in seven federal states, including Veracruz,

Jalisco and Chihuahua. Specifically, Veracruz contributes 13% to the national livestock production, with a herd of 5.8 million heads in various systems ranging from highly technified feedlots to extensive grazing operations (Almeraya Soberanes *et al.*, 2024; Espinoza-Cantú *et al.*, 2026). Intensive production enhances growth performance but increases the risk of disease and antimicrobial use (Rossi *et al.*, 2022). BRDC is particularly impactful in this context (Adekunle *et al.*, 2026) and often presents subclinical cases that remain undetected during production. For every clinically affected animal, up to four additional animals may have significant pulmonary lesions without observable signs of the disease (Kamel *et al.*, 2024). This highlights the limitations of traditional diagnostic approaches.

Artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL) (Gupta *et al.*, 2022; Kufel *et al.*, 2023; Sharifani & Amini, 2023), has emerged as a valuable tool for analyzing animal health data and supporting diagnostic tasks (Bumbálek *et al.*, 2025; Guarnido-Lopez *et al.*, 2024; Xiao *et al.*, 2025; Yang *et al.*, 2025). In this context, computer vision models enable the automated analysis of veterinary images. Deep learning architectures allow the identification and segmentation of anatomical structures and lesions, facilitating the objective quantification of lung damage and reducing diagnostic subjectivity.

Segmentation, defined as the division of an image into meaningful regions, is a key computer vision technique that enables the identification of relevant structures for analysis. It has been successfully applied in medical imaging (Houssein *et al.*, 2021), mineral analysis (Liu *et al.*, 2021), and veterinary sciences, including the segmentation of cattle body parts and animal activity recognition using DL approaches (Bello *et al.*, 2024; Jia *et al.*, 2021). In related studies, automated models have been applied to lung image analysis in cattle and pigs, showing promising but still limited performance, particularly in lesion detection (Bortoluzzi *et al.*, 2023; Recchia *et al.*, 2025).

This study aimed to assess the potential of modern deep learning algorithms as a decision-support tool for post-mortem veterinary evaluations. The results demonstrate promising performance and support the feasibility of these approaches as assistive technologies, while also highlighting opportunities for further refinement and validation using larger and more representative datasets to strengthen their practical implementation. Future studies should include substantially larger and more heterogeneous datasets, ideally collected from multiple facilities and production systems, to improve model robustness, assess external validity, and confirm the reproducibility of the reported results.

MATERIALS AND METHODS

The images were captured under consistent conditions using a Nikon D5500 camera at a fixed distance from an aluminum base where the lungs were placed, resulting in images of 4000 × 6000 pixels from a single slaughter-

house. The slaughter line personnel separated the lungs from other organs and tissues before placing them on the base for photography. Each lung was photographed twice: once to capture the dorsal view and again after being turned over to capture the ventral view.

This study was approved by the Bioethics Committee of the School of Veterinary Medicine and Animal Sciences at Veracruz University (approval registration number CO-BIBA011/2021).

Before initiating AI training, four veterinary annotators from the Universidad Veracruzana (UV) performed manual segmentation, and inter-annotator agreement was evaluated. The initial segmented areas included the lungs, cuts, trachea and extra tissues, as well as findings or lesions. Findings or lesions were differentiated from healthy tissue based on a systematic evaluation of the pulmonary parenchyma (the functional, gas-exchanging tissue of the lungs, comprising respiratory bronchioles, alveolar ducts, and alveoli), specifically focusing on color alterations, morphological changes, and surface irregularities. These annotations were later validated by a single veterinary pathology expert.

The resulting annotations were stored in JSON format (*JavaScript Object Notation*), which required conversion into a format compatible with the DL training framework. Automatic preprocessing algorithms were employed to facilitate this conversion from JSON to the required structure, generating a YAML (*Yet Another Markup Language*) configuration file containing all necessary metadata for training and that can be manually edited to modify the dataset path. To enhance the results, the data were prepared for k-fold cross-validation. Specifically, five different sets of images were created from 129 images divided into an 80-20 ratio (103 for training and 26 for validation), ensuring that every image was used for both training and validation at least once. This technique provides a robust and reliable estimate of the model performance on unseen data by optimizing data utilization.

Through repeated training and validation across multiple data partitions, this approach minimizes dependence on a single split, reduces the variance in performance estimates, and improves the assessment of the model's ability to generalize.

Training algorithm

Hardware Specifications

Training was conducted on a workstation equipped with an NVIDIA GeForce RTX 4060 (8GB VRAM), a 13th Gen Intel Core i7-13650HX CPU, and 16GB RAM, running Windows 11. The models were implemented in Visual Studio Code (VS Code) using Python 3.12, with PyTorch 2.10 and CUDA 12.9 (Compute Unified Device Architecture).

Training Parameters

Several training sessions were conducted to develop algorithms capable of identifying and segmenting bo-

vine lungs into a) lung with trachea and extra tissue, and b) findings or lesions. Separate identification and segmentation were performed to ensure clear visualization and distinction between the lung regions and pathological findings, as both overlapped within the same area of interest.

The training algorithm was executed multiple times, first without data augmentation, and subsequently using various data augmentation techniques described by Ultralytics (n.d.-a). The model was trained using the YOLO11n-seg architecture (2.9 million parameters) configured with the Adam optimizer, an input image size of 640 pixels, a learning rate of 0.001, and a batch size of 4. Early stopping was set with a patience of 5 epochs, using 1 worker and a fixed random seed of 1801. The overall training procedure is summarized in Table 1.

The model was trained five times using the YOLO11n-seg.pt pretrained instance segmentation model from the Ultralytics YOLO family. The first fold of the segmented images was used for the initial training, and

the resulting best.pt file training was applied to each of the subsequent folds of images in turn for further training.

To evaluate the object detection performance, several metrics were considered (Trigka & Dritsas, 2025; Ultralytics, n.d.-b). These metrics assess the effectiveness of a model in identifying and localizing objects within images. The metrics used are shown and explained in Table 2.

These metrics provide a comprehensive evaluation of the model’s performance in identifying and segmenting bovine lungs and pathological findings.

Segmentation and pixel overlay

Using images automatically segmented by the ML program, we developed another algorithm (Algorithm 1) to produce masks for pixels identified as lungs and findings. The segmentation procedure marks the contours of the lungs and findings in the original image. Subsequently, the areas corresponding to both regions were computed and used to calculate the percentages of healthy and affected lung regions.

Table 1. Summary of training procedures, including applied data augmentation techniques. The training utilized standard data augmentation methods to improve model generalization and reduce over-fitting. These training sessions were repeated 5 times each fold for k-fold cross-validation.

Sections	Epochs	Batch	Patience	Model name	Data augmentation configuration
Lung and trachea	100	4	5	Train 0	None
				Train 1	hsv_s=0.5, hsv_v=0.5
				Train 2	hsv_s=0.5, hsv_v=0.5, translate=0.15, scale=0.25, shear=5
				Train 3	hsv_s=0.5, hsv_v=0.5, translate=0.15, scale=0.25, shear=5, flipud=1.0, fliplr=1.0, erasing=0.6
Findings	100	4	5	Train H0	None
				Train H1	hsv_s=0.5, hsv_v=0.5
				Train H2	hsv_s=0.5, hsv_v=0.5, translate=0.15, scale=0.25, shear=5
				Train H3	hsv_s=0.5, hsv_v=0.5, translate=0.15, scale=0.25, shear=5, flipud=1.0, fliplr=1.0, erasing=0.6

The data augmentation techniques (Ultralytics, n.d.-a) used were as follows: erasing, randomly erases parts of the image during classification training; flipud (flip up-down), vertically flips the image by reflecting it along the x-axis; fliplr (flip left-right), horizontally flips the image by reflecting it along the y-axis; hsv (hue, saturation, value), a color representation model in which hue defines the color, saturation defines its intensity, and value defines its brightness; hsv_s, modifies the saturation (S) component, changing the intensity or vividness of colors; hsv_v, modifies the value (V) component, adjusting overall brightness; scale, randomly resizes images within a specified range; shear, skews images along the x and y axes, effectively shifting parts of the image in one direction while maintaining parallel lines; and translate, shifts images horizontally and vertically by a random fraction of the image size.

Table 2.
Summary of the metrics used to assess the effectiveness of the resulting models.

Metric	Purpose/Definition	Formula/benchmark
Specificity or Precision	Measures the proportion of identified objects that are correct.	
Sensitivity or Recall	Assesses the number of real objects correctly identified by the model. High-sensitivity models often achieve 0.85-0.90+ in studies, particularly with large lesions.	
F1 score	Evaluates accuracy by calculating the harmonic mean of precision and recall. It balances precision (how many predicted positives are correct) and recall (how many actual positives were found).	$F1=2\times(Precision\times Recall)/(Precision+Recall)$ Or $F1=2\times TP/(2TP+FP+FN)$ Where: TP = True positive FP = False positive FN = False negative
F2 score	A classification metric that measures the model performance by calculating a weighted harmonic mean of precision and recall, with more emphasis on recall. Used when minimizing false negatives is critical.	$F2=5\times(Precision\times Recall)/(4\times Precision+Recall)$
Mean Average Precision (mAP)	Evaluates object detection model performance. mAP provides a single, comprehensive score that summarizes a model's ability to correctly identify and localize objects across various classes and confidence levels.	Excellent Performance: mAP > 60% Good Performance: mAP 50% - 60% Medium Performance: mAP 40% - 50% Poor Performance: mAP < 40%
Losses – Segmentation loss	The loss function for segmentation quantifies the discrepancy between the predicted segmentation mask and ground truth annotations. Lower values indicate better agreement.	
Losses – Classification Loss	The loss function for object classification quantifies the difference between the predicted class probabilities and true class labels. Lower values indicate more accurate class predictions.	

RESULTS

Lung and trachea identification

The first objective was to automatically identify the lungs and trachea using artificial intelligence; therefore, the training process was initiated to enable their automatic differentiation from each other. The average metrics of each set of five trainings across the four different training configurations (one without data augmentation and three with different data augmentations, as shown in Table 1) are presented in Table 3.

The results in Table 3 show clear differences in the final average performance values across various training runs. Notably, Train 2 exhibited the best overall performance, achieving a precision of 0.869, recall of 0.804, F1 score of 0.827, and F2 score of 0.806. This model also achieved the highest mAP@50 value (0.842), indicating superior object detection and localization capabilities. Additionally, Train 2 showed lower values for the loss functions, suggesting a more effective convergence during training and validation.

Trains 0 and 1 showed competitive but inferior performance compared to Train 2. Train 0 achieved the highest precision (0.874) but with a lower recall (0.790), indicating

a slight tendency to produce fewer false positives at the expense of missing some detections. In contrast, Train 1 presented a different balance between accuracy and recall but did not surpass the overall performance of Train 2. Both models had mAP@50 values close to Train 2 (0.838 and 0.836, respectively), although with slightly higher losses, reflecting less optimal convergence.

In contrast, Train 3 shows significant performance deterioration in all evaluated metrics, with considerably lower precision (0.482), recall (0.434), F1-score (0.427), F2-score (0.415), and mAP@50 (0.392). Despite employing more aggressive data augmentation techniques, these results suggest that excessive transformations can negatively impact model learning, hindering correct identification, segmentation, and classification. Overall, the intermediate configuration (Train 2) offers the best compromise between robustness and generalizability, demonstrating superior performance and effective convergence.

Figure 1 illustrates the evolution of the performance metrics and loss functions over epochs for the Train 2 configuration, which exhibited the best overall performance among the four training settings (Table 3). These results are based on the final training with the fifth cross-validation set for Train 2.

Table 3.
Average performance metrics of the trained models.

	Epochs	Precision (M)	Recall (M)	F1 score (M)	Segmentation loss (Train / Val)
Train 0	9.62	0.874*	0.790	0.820	1.187 / 2.261
Train 1	9.34	0.857	0.793	0.812	1.171 / 2.351
Train 2	11.05*	0.869	0.804*	0.827*	1.096* / 2.152*
Train 3	6.63	0.482	0.434	0.427	1.109 / 6.992
		F2 score (M)	mAP 50 (M)	mAP 50-95 (M)	Class loss (Train/Val)
Train 0		0.793	0.838	0.678*	0.988 / 1.102
Train 1		0.789	0.836	0.675	0.980 / 1.067
Train 2		0.806*	0.842*	0.673	0.969* / 0.998*
Train 3		0.415	0.392	0.161	0.993 / 2.448

Train 0: no augmentation; Train 1: hsv_s and hsv_v; Train 2: hsv_s, hsv_v, translate, scale, and shear; Train 3: hsv_s, hsv_v, translate, scale, shear, flipud, fliplr, and erasing.
M: Segmentation mask (indicates metrics evaluated on mask-level predictions).
*Indicates the best result among the four configurations. For Segmentation loss and Class loss, values are reported as training/validation.

For precision, recall, F1-score, and F2-score metrics, a sharp increase occurred during the first epochs, rising from low values in the initial epoch to values higher than 0.90 by the second epoch. These results suggest rapid learning of relevant discriminative features, good detection capability without significant sensitivity losses, and robust and balanced model performance.

Regarding the loss functions, the segmentation loss showed a drastic decrease in the early epochs, dropping from high values to values close to 1.0, and then remaining stable with small fluctuations. This suggests a controlled optimization process. The classification loss showed a general decreasing trend, although with slight variations throughout training. A significant initial reduction was fol-

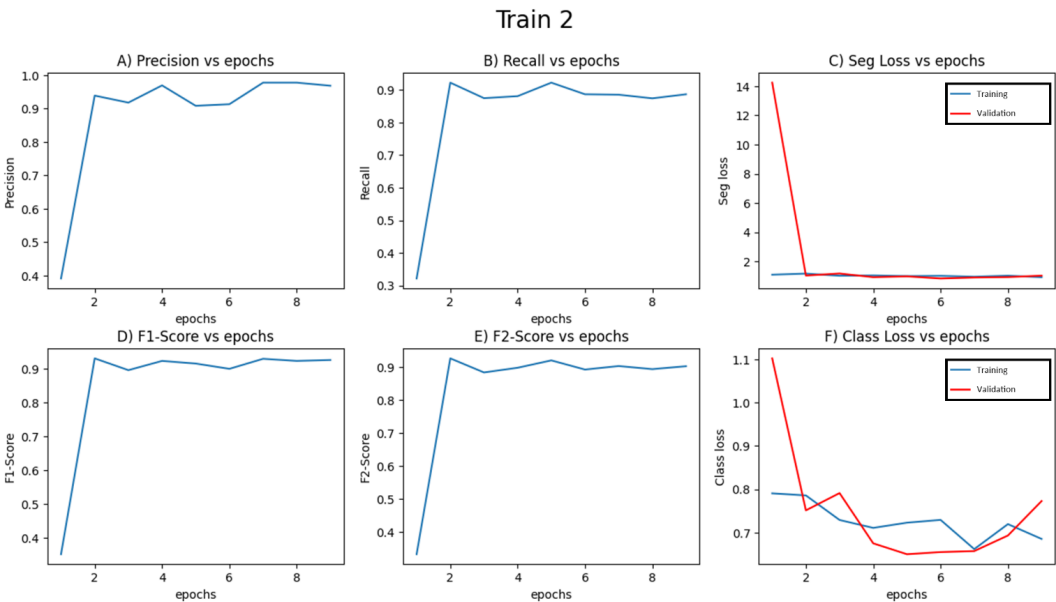


Figure 1.
Learning curves for Train 2 from the fifth cross-validation fold. Precision (A), recall (B), F1-score (D), and F2-score (E) increased rapidly during initial epochs before stabilizing, reaching final values of 0.968, 0.886, 0.925, and 0.902, respectively. Training and validation segmentation losses (C) converged to 0.914 and 1.020, while classification losses (F) reached 0.684 and 0.772, indicating stable convergence and strong generalization.

lowed by small fluctuations that remained within a narrow range. This implies that the model reached a good convergence point and continued to fine-tune its parameters.

Overall, the results demonstrate that Train 2 achieved rapid and stable convergence after previous fold trainings, with a suitable balance between precision and detection capability. Metric stability and sustained loss reduction suggest good generalizability and the absence of significant overfitting.

Figure 2 illustrates the area under the precision-recall curve (AUPRC) for the final training with the fifth cross-validation set. These curves were used because of the dataset imbalance between healthy and unhealthy lungs and to evaluate how well the model balances precision and recall across various confidence thresholds.

In comparison, Figure 2A (Train 0), 2B (Train 1), and 2C (Train 2) show consistent and robust behaviors, with

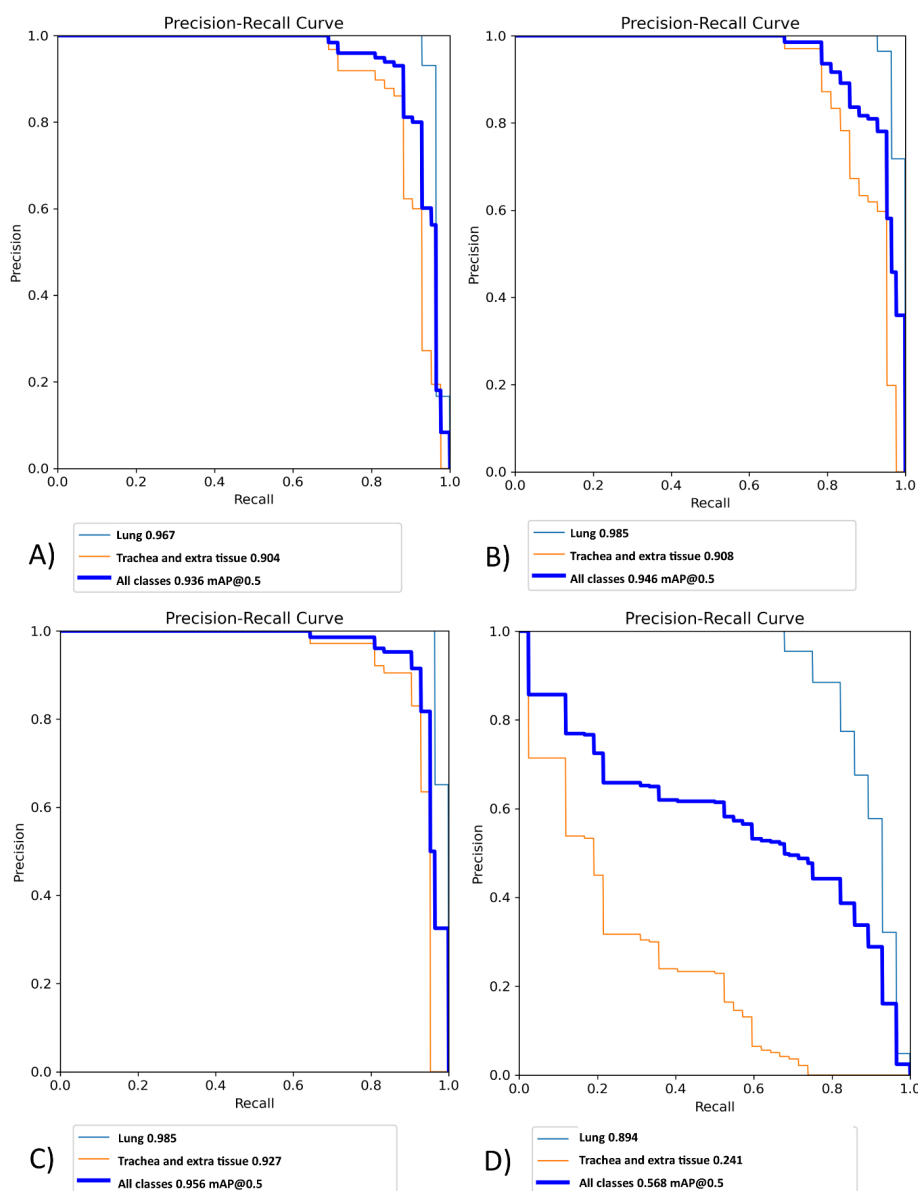


Figure 2.

Precision-Recall curves (AUPRC) from the fifth cross-validation fold for the anatomical segmentation models. (A) Train 0 achieved AUPRCs of 0.967 for lung and 0.904 for trachea and extra tissue, yielding an overall mAP@0.5 of 0.936. (B) Train 1 reached AUPRCs of 0.985 (lung) and 0.908 (trachea and extra tissue), with an mAP@0.5 of 0.946. (C) Train 2 achieved the best overall performance, obtaining AUPRCs of 0.985 (lung) and 0.927 (trachea and extra tissue) for an mAP@0.5 of 0.956. (D) Train 3 showed lower performance, with AUPRCs of 0.894 (lung) and 0.241 (trachea and extra tissue), resulting in an mAP@0.5 of 0.568.

curves that remain close to the upper right vertex of the precision-recall space, indicating a high overall performance. The *lung* class exhibits particularly high AUPRC values (0.967 for Train 0, 0.985 for Train 1, and 0.985 for Train 2), demonstrating the model's excellent ability to identify the lung structure with high precision and sensitivity. The *trachea and extra tissue* class exhibits slightly lower but still competitive performance, with AUPRC values of 0.904 (Train 0), 0.908 (Train 1), and reaching its best result in Train 2 at 0.927. Overall, the aggregate model performance improved incrementally, increasing from a mAP@0.5 of 0.936 (Train 0) to 0.946 (Train 1) and reaching its best value in Train 2 at 0.956. These results are consistent with the previously analyzed metrics, confirming that Train 2 offers the best balance between accuracy and recall. In contrast, Figure 2D (Train 3) shows a significantly poorer performance, with the *lung* class at 0.894 and the *trachea and extra tissue* class at 0.241. The curve shape in this case demonstrates a rapid drop in accuracy as recall increases, indicating a considerable increase in the number of false positives. Overall, Train 2 was the most suitable

configuration and achieved the best performance. Finally, the resulting training images were used on dorsal and ventral images. An example of the model's classification and segmentation results is shown in Figure 3.

Figure 3A and 3B serve as the ground truth for comparing the model's classification. In Figure 3A0 and 3B0 (Train 0), the model adequately captured the general lung shape, although it exhibited some inaccuracies at the edges and in delineating more complex structures. Additional data for Train 1 (Figure 3A1 and 3B1) slightly improved the segmentation consistency, resulting in fewer errors in the peripheral regions and better adaptation to texture variations. The results of Train 2 (Figure 3A2 and 3B2) show the greatest visual agreement with the reference annotations. The model achieves more precise contour delineation, especially in anatomically complex areas, and maintains more uniform segmentation throughout the lung structure. In contrast, the results of Train 3 (Figure 3A3 and 3B3) show low classification and segmentation performance. Most regions were misclassified as background, with a loss of mask continuity and minimal proper segmentation.

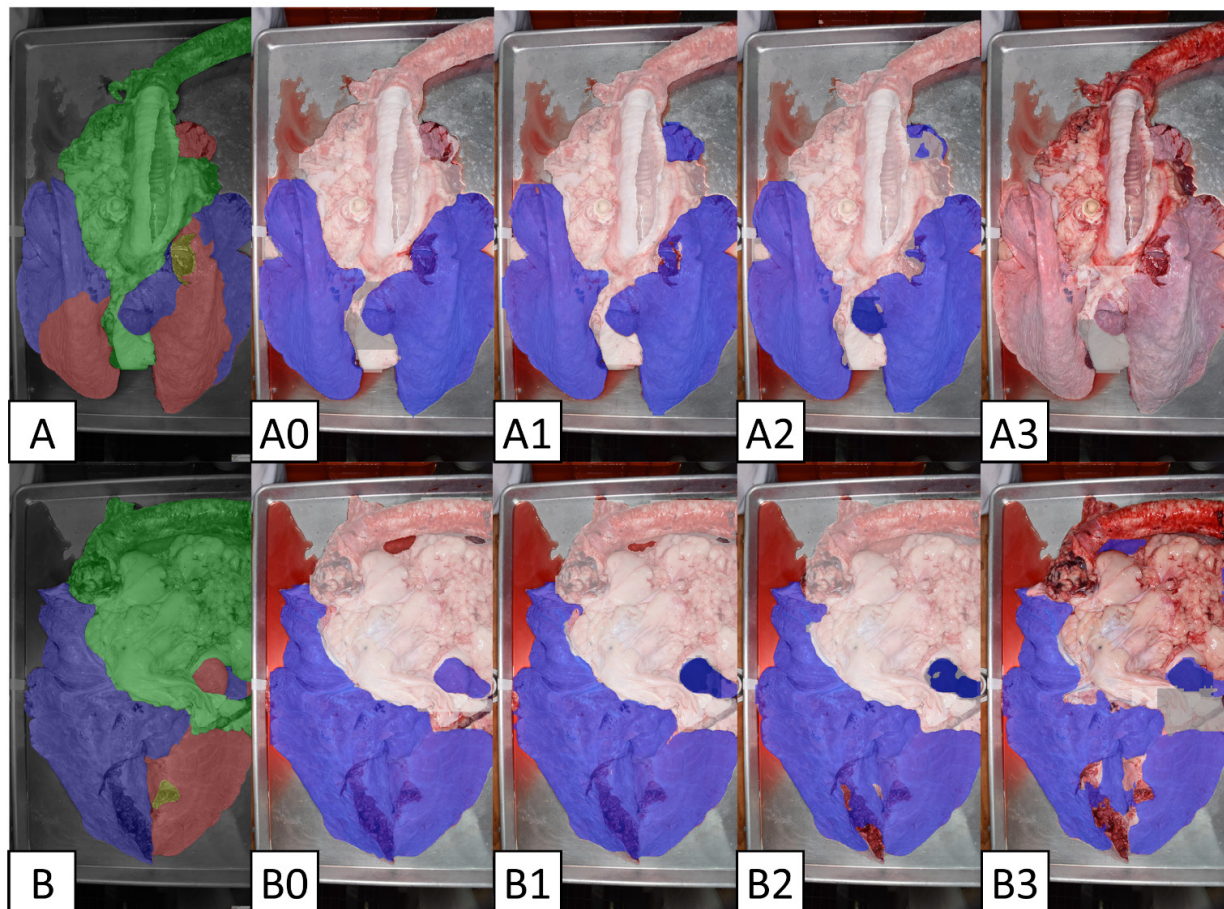


Figure 3.

Qualitative comparison of segmentation results across the four training configurations. Ground truth annotations for the dorsal (A) and ventral (B) views highlight healthy lung tissue (red), affected tissue/findings (blue), and trachea with extra tissues (green). Panels A0–A3 and B0–B3 present the predictions generated by the final models from the fifth cross-validation fold for Train 0 to Train 3, respectively.

Notably, training 2 provided the most accurate and consistent lung segmentation, further supporting its overall superior performance.

Automatic finding identification

The next objective was to automatically identify findings using DL. Similar to lung and trachea segmentation, we conducted separate trainings with distinct data augmentation strategies (Table 1) along with five rounds of k-fold cross-validation. Table 4 presents the results for the Findings task under different training configurations (Train H0–H3). The overall performance was lower than that of the Lung and Trachea tasks, suggesting greater detection complexity.

Table 4. Average performance metrics of the trained models.

	Epochs	Precision (M)	Recall (M)	F1 score (M)	Segmentation loss (Train / Val)
Train H0	8.78	0.661	0.544	0.593	2.070 / 3.502
Train H1	11.01*	0.650	0.583*	0.602*	2.024 / 3.166
Train H2	8.08	0.663*	0.565	0.597	2.035 / 3.063*
Train H3	7.74	0.383	0.363	0.354	1.954* / 5.288
		F2 score (M)	mAP 50 (M)	mAP 50-95 (M)	Class loss (Train / Val)
Train H0		0.560	0.532	0.292	1.524 / 1.679
Train H1		0.579*	0.572*	0.323*	1.451* / 1.598*
Train H2		0.566	0.542	0.298	1.506 / 1.631
Train H3		0.348	0.284	0.117	1.462 / 2.304

Train H0: no augmentation; Train H1: hsv_s and hsv_v; Train H2: hsv_s, hsv_v, translate, scale, and shear; Train H3: hsv_s, hsv_v, translate, scale, shear, flipud, fliplr, and erasing.
M: Segmentation mask (indicates metrics evaluated on mask-level predictions).
* Indicates the best performance. Segmentation loss and Class loss are reported as training/validation values.

Train H1 exhibited the best overall performance, achieving the highest recall (0.583), F1-score (0.602), F2-score (0.579), mAP@50 (0.572), and mAP@50–95 (0.323). This indicates a better ability to correctly detect regions of interest while maintaining a suitable balance between accuracy and sensitivity. Train H1 also exhibits more stable convergence with the lowest classification losses (1.451 and 1.598). Train H0 and Train H2 exhibit similar performances. Train H0 achieved a precision of 0.661, whereas Train H2 obtained the highest precision value (0.663); however, both showed lower recall and other metrics than Train H1, indicating a reduced ability to recover all relevant instances. In contrast, Train H3 showed the worst performance in all the evaluated metrics. Precision (0.383), recall (0.363), F1-score (0.354) and F2-score (0.348) decrease considerably, accompanied by a sharp drop in mAP@50 (0.284) and mAP@50–95 (0.117). Interestingly, Train H3 had a slightly

lower segmentation loss (1.954) than the other models but showed a significant increase in the validation segmentation loss (up to 5.288). Additionally, its training classification loss was relatively low (1.462), although it increased notably during validation (up to 2.304), indicating difficulties in correctly segmenting and identifying classes.

Figure 4 shows the performance metrics and loss curves over epochs for Train H1, which exhibited the best overall performance (Table 4). These results illustrate the final training in the fifth cross-validation fold. Precision (Figure 4A), recall (4B), F1-score (4D), and F2-score (4E) increase sharply during the initial epochs before exhibiting inconsistent behavior, fluctuating considerably until the end of the training. In most cases, the metrics stabilize at values

above 0.70. These results suggest a rapid initial learning, followed by an unstable optimization phase with significant performance oscillations, indicating partial convergence with acceptable final outcomes.
Regarding the loss functions (Figures 4C and 4F), both the training and validation segmentation losses exhibit a consistent downward trend over epochs. The training segmentation loss curve (Figure 4C) shows that the model progressively learns to better delineate the regions of interest, despite some intermediate oscillations. Similarly, the validation loss decreases more prominently during the early epochs before stabilizing with minor fluctuations. This suggests effective learning during training and significant initial improvement in generalization, which remained relatively stable in later epochs.
On the other hand, training classification loss gradually decreases, indicating that the model improves its ability

Train H 1

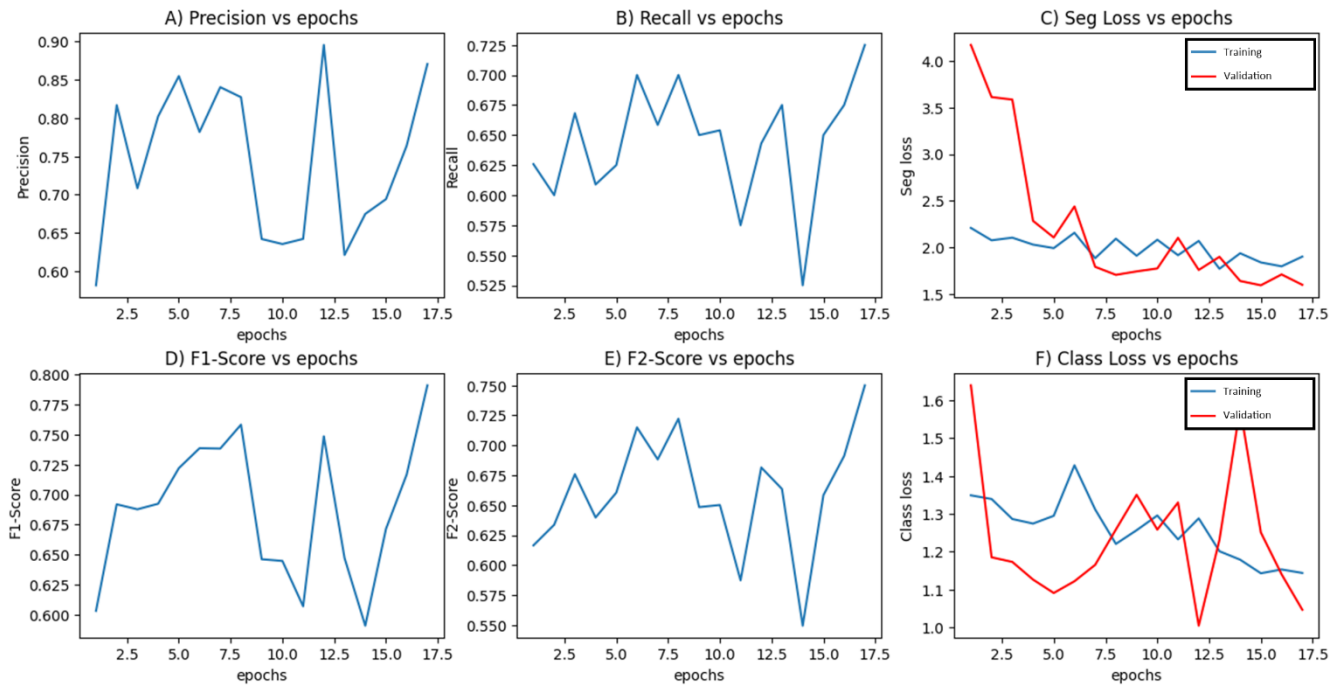


Figure 4.

Learning curves for Train H1 from the fifth fold of the 5-fold cross-validation. Precision (A), recall (B), F1-score (D), and F2-score (E) exhibited greater variability throughout training than the anatomical segmentation models, reaching final values of 0.870, 0.725, 0.791, and 0.750, respectively. Training and validation segmentation losses (C) decreased progressively to 1.899 and 1.596, while classification losses (F) reached 1.143 and 1.046, respectively. Despite fluctuations, both loss functions showed an overall downward trend, indicating model convergence.

to correctly assign class labels despite slight variations. However, validation classification loss is less stable -it decreases initially but then exhibits intermittent increases, suggesting that the model struggles to achieve full classification generalization. Overall, results demonstrate effective, though imperfect, learning. The findings indicate that the model rapidly learns meaningful features during the initial training stages. However, the evaluation metrics and loss values suggest that the optimization process is not fully stable. These oscillations may indicate limited consistency and difficulty in achieving complete classification generalization.

Figure 5 illustrates the area under the precision-recall curve (AUPRC) for training to differentiate between healthy and unhealthy tissue. Given the dataset imbalance due to a lack of completely healthy lungs, it is crucial to determine whether the model can correctly identify the difference between healthy tissue and findings.

The Precision-Recall curves illustrate the varying ability of the models to maintain high precision as recall increases. Train H1 (Figure 5B) exhibited the best overall performance, achieving an mAP@0.5 of 0.803. Its curve maintains high precision across a broad recall range, reflecting a solid balance between false positives and false negatives, which indicates high robustness. Train H2 (Figure 5C) followed with

an mAP@0.5 of 0.735, closely trailed by Train H0 (Figure 5A) at 0.727. Both models displayed similar profiles: high precision at low-to-medium recall, followed by steeper declines at higher recall values, showing that their performance decreases when retrieving all instances. In contrast, Train H3 (Figure 5D) showed the lowest performance (mAP@0.5 = 0.520), with a curve that drops earlier and more steadily, reflecting reduced generalizability and higher error rates. Together, these results demonstrate that Train H1 was the most effective model for detection.

Next, as with the previous lung identification experiments, we applied the resulting models to dorsal and ventral images (Figure 6). Similar to the layout of Figure 3, panels 6A and 6B serve as the ground truth for comparing classification results.

In the dorsal view (Figure 6A), Train H1 (Figure 6A1) shows the closest visual match with the ground truth annotation, achieving adequate coverage of affected lung regions and precise border delineation. Train H2 (Figure 6A2) demonstrates similar performance but with slight overestimation in some areas, while Train H0 (Figure 6A0) tends to undersegment regions and exhibits inconsistent coverage. On the other hand, Train H3 (Figure 6A3) produces more diffuse and less precise segmentations, erroneously including non-target areas.

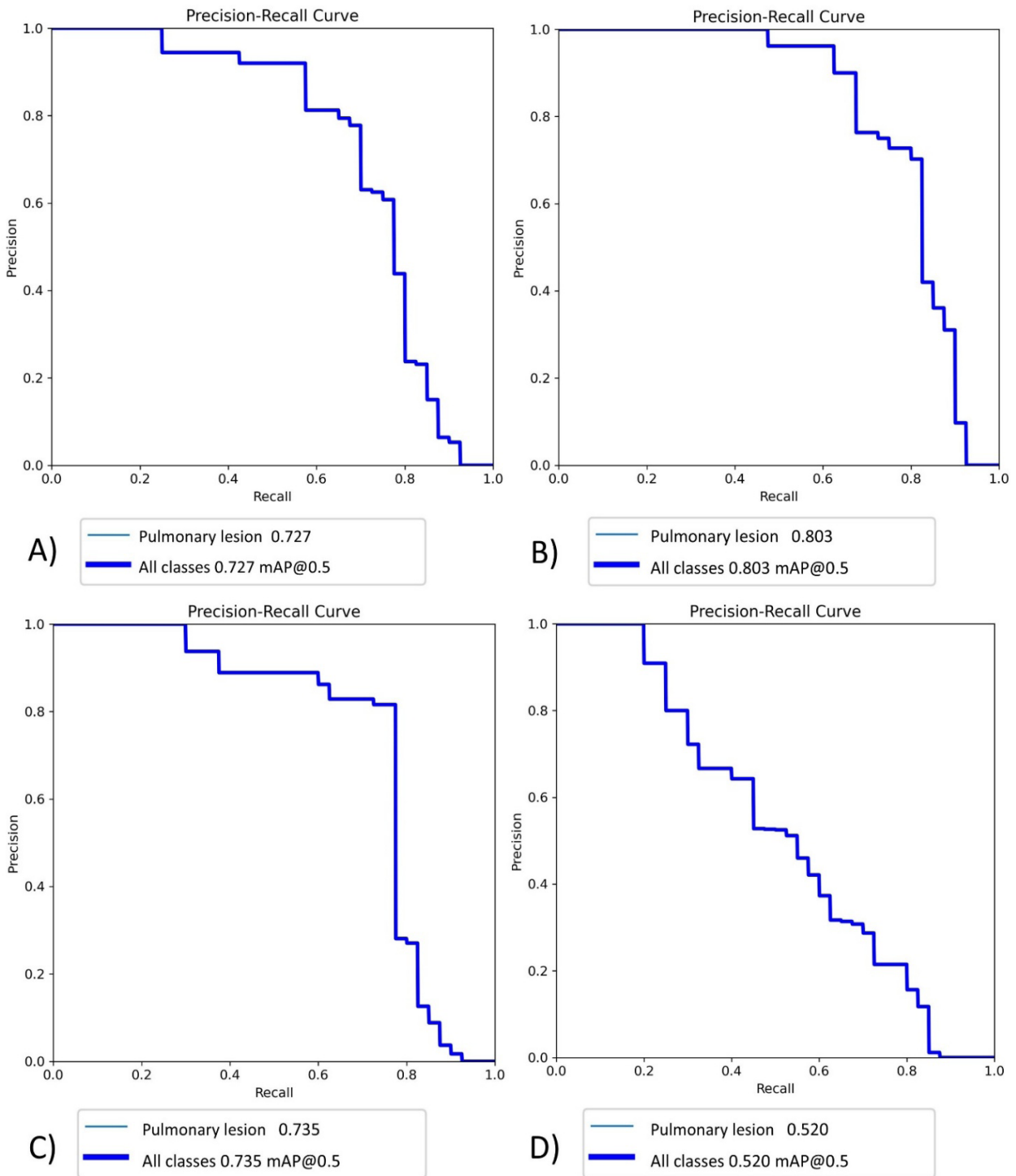


Figure 5.

Precision-Recall curves (AUPRC/mAP@0.5) from the fifth cross-validation fold for the pulmonary lesion detection models. (A) Train H0 achieved an AUPRC of 0.727. (B) Train H1 achieved the best performance with an AUPRC of 0.803. (C) Train H2 reached an AUPRC of 0.735. (D) Train H3 showed a marked performance reduction, obtaining an AUPRC of 0.520.

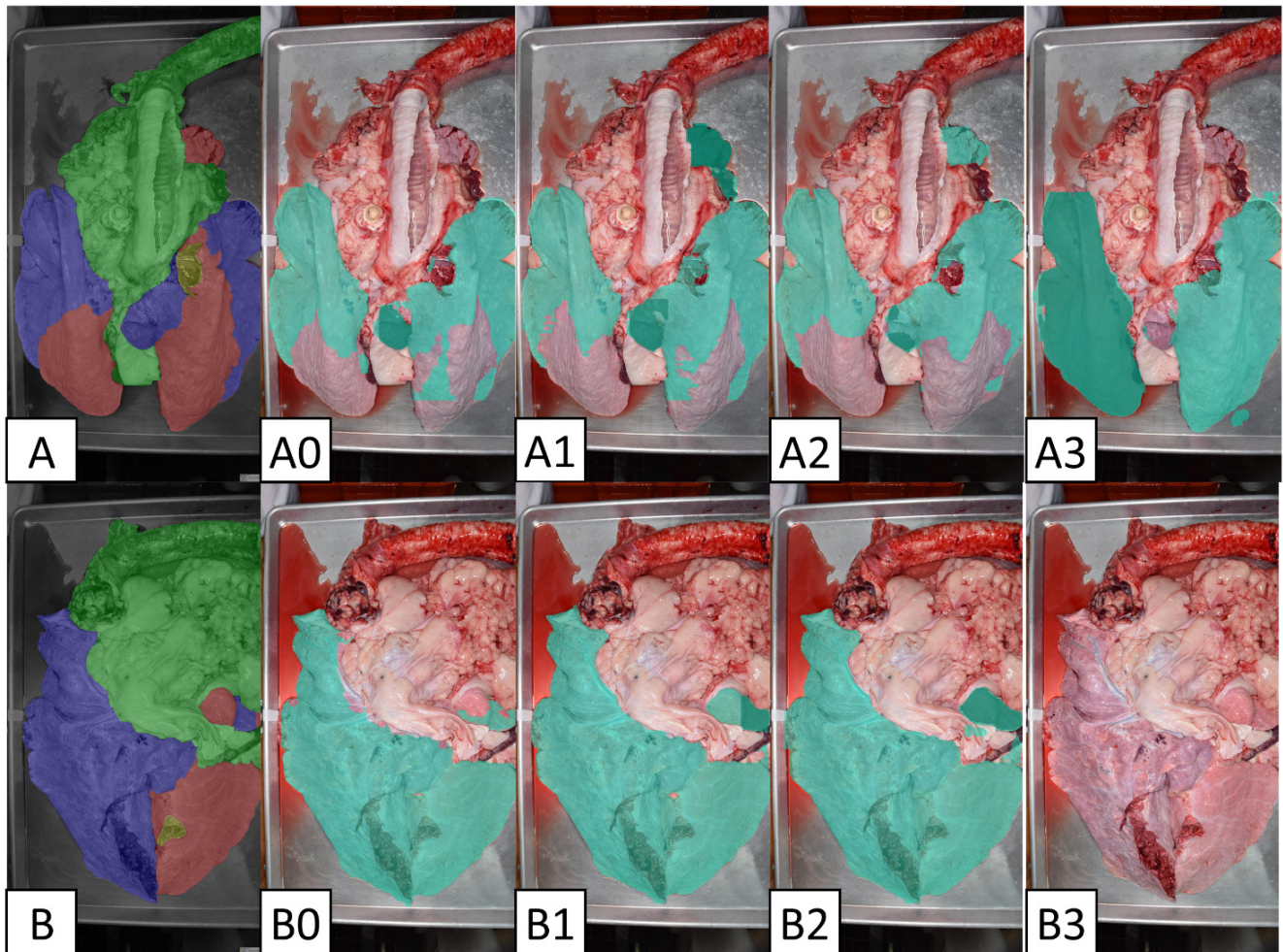


Figure 6.

Qualitative comparison of segmentation results for the four training configurations. Panel A corresponds to the dorsal view and Panel B to the ventral view. A and B show the ground truth annotations (findings in blue), while A0–A3 and B0–B3 present the predictions generated by the final models obtained from the fifth fold of the 5-fold cross-validation for Train H 0–3, respectively, with cyan indicating the regions identified as lesions or findings by the models.

A consistent trend is observed in the ventral view (Figure 6B). Train H1 (Figure 6B1) again achieves segmentation closest to the ground truth, maintaining a strong balance between coverage and accuracy. Train H2 (Figure 6B2) remains competitive but incorporates more irrelevant tissue, whereas Train H0 (Figure 6B0) displays limitations in region extent. Train H3 (Figure 6B3) exhibits the poorest fit, struggling to identify, classify, and segment findings. Overall, these qualitative observations support the quantitative findings: Train H1 offers the most accurate and consistent segmentation, followed by Train H2 and Train H0, while Train H3 shows inferior performance.

Segmentation and pixel overlay

Figure 7 illustrates an example of the automatic contouring produced from model results and calculated areas

for healthy and unhealthy (where findings were found) tissue after applying Algorithm 1 (Table 4). Figure 7A displays the original post-mortem image, providing a baseline view without any segmentation or annotation. Figure 7B shows automated delineation of lung regions (blue contours) and lesion areas (red contours) generated through the post-processing pipeline. By segmenting regions enclosed within these contours and using them to generate masks and calculate the corresponding pixel areas, we obtained the results shown in Figure 7C, including the percentage distribution of “healthy” and “unhealthy” tissue.

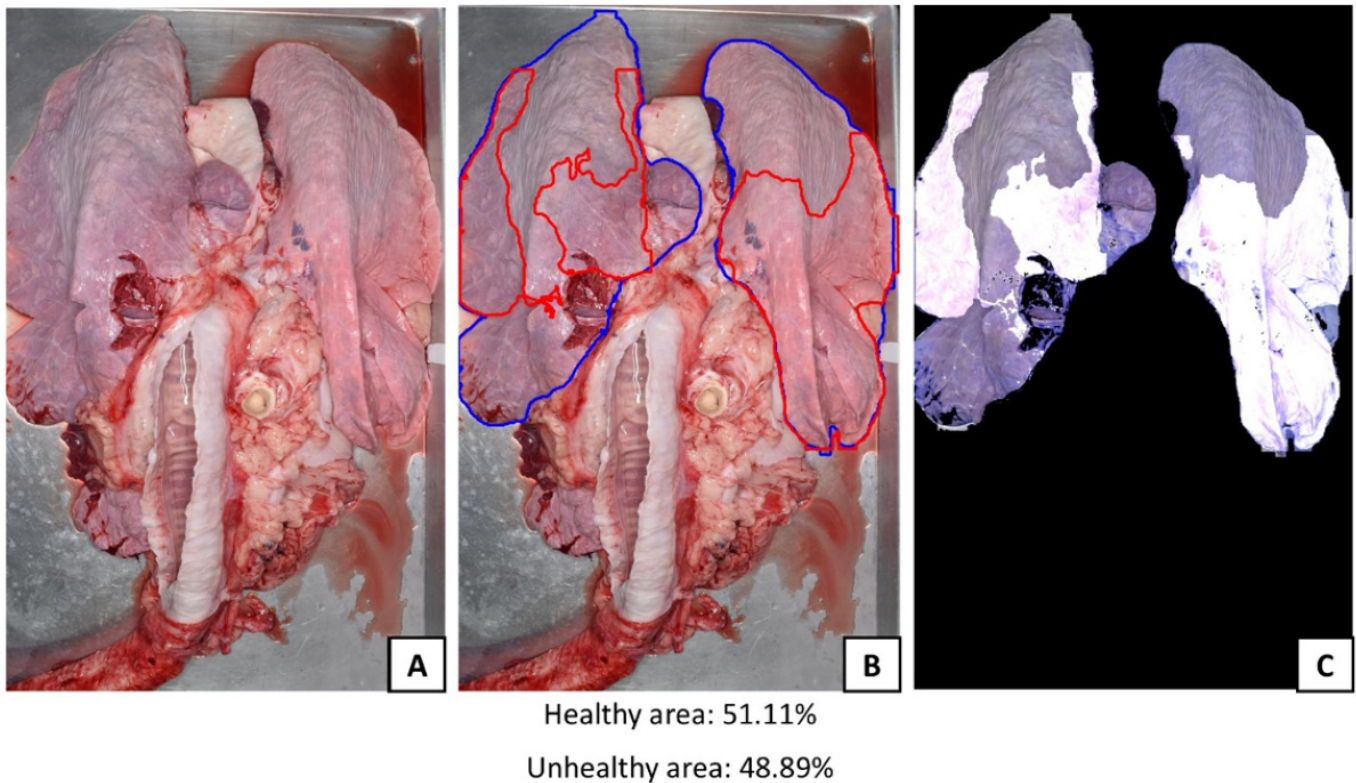


Figure 7.

Example of the proposed automated lung health assessment workflow. (A) Original post-mortem image. (B) Automatic delineation of the lung (blue) and lesion (red) contours overlaid on the original image. (C) Pixel-level overlap between the segmented lung area and detected lesions, from which healthy (51.11%) and unhealthy (48.89%) tissue percentages are automatically computed.

Algorithm 1.

Pseudocode for calculating healthy and unhealthy lung tissue percentages

Input:

Original lung image
ML-segmented lung image
ML-segmented finding image

Output:

Contour-marked lung image
Healthy tissue percentage
Unhealthy tissue percentage

1. Load images

Load the original lung image.
Load the ML-segmented lung image.
Load the ML-segmented finding image.

2. Generate masks

Generate a binary mask for the lung region.
Generate a binary mask for the finding region.

3. Determine total lung area

Convert the lung mask to grayscale.
Apply threshold truncation using a threshold of 127 and a maximum value of 255.
Detect external contours.
For each contour:
Calculate its perimeter.

If the perimeter is greater than 5000 pixels:
Retain the contour.
Calculate its area.
Sum the areas of all retained contours to obtain LungArea.

4. *Determine unhealthy tissue area*

Convert the finding mask to grayscale.
Apply the same thresholding procedure.
Detect external contours.
For each contour:
Calculate its perimeter.
If the perimeter is greater than 5000 pixels:
Retain the contour.
Calculate its area.
Sum the areas of all retained contours to obtain FindingArea.

5. *Calculate healthy tissue area*

HealthyArea = LungArea – FindingArea

6. *Calculate tissue percentages*

If LungArea > 0:

HealthyPercentage = (HealthyArea / LungArea) × 100

UnhealthyPercentage = (FindingArea / LungArea) × 100

Otherwise, set both percentages to 0.

7. *Generate output*

Overlay the retained lung and finding contours on the original image.
Save the contour-marked image.
Report the healthy and unhealthy tissue percentages.

As shown in Figure 7C, the algorithm successfully quantified the lesion burden. In this representative case, 51.11% of the segmented lung area was classified as healthy, while the remaining 48.89% corresponded to unhealthy tissue affected by lesions.

DISCUSSION

Anatomical pulmonary segmentation (lung vs trachea and extra tissues)

The strong performance achieved in anatomical segmentation was expected, given the well-defined boundaries and consistent appearance of the lung and trachea across images. Unlike pathological lesions, these anatomical structures exhibit lower variability, making them more suitable targets for deep learning models trained on limited datasets. Nevertheless, dataset size remains a limitation. Although data augmentation and cross-validation were employed to enhance robustness, the observed performance may not fully translate to independent datasets collected under different slaughterhouse conditions. Consequently, these results should be interpreted as proof of feasibility rather than definitive evidence of generalizable performance.

Previous research demonstrates that anatomical structures with consistent morphology are generally segmented more accurately than pathological regions, particularly when deep learning is applied to livestock imaging (Jia et

al., 2021; Bello et al., 2024). Furthermore, recent reviews recognize anatomical segmentation as one of the most mature applications of computer vision in precision livestock farming (Guarnido-Lopez et al., 2024; Yang et al., 2025).

A notable contribution of this study is the demonstration that reliable anatomical segmentation can be achieved from post-mortem slaughterhouse images in the same production line, which typically present greater variability in lighting, positioning, contamination, and tissue presentation than images acquired under controlled laboratory conditions. Furthermore, unlike other studies that use only a single lung view (Bortoluzzi et al., 2023; Recchia et al., 2025), this study incorporated images from both the dorsal and ventral sides, allowing the evaluation of lesions across complementary anatomical regions and supporting a more complete characterization of pulmonary pathology. This dual-view strategy closely resembles routine veterinary post-mortem assessments and enhances the practical relevance of the proposed framework.

Pulmonary lesion detection

Compared to anatomical segmentation, pulmonary lesion detection presents a greater challenge—compounded by dataset limitations—due to lesion heterogeneity, subtle pathological features, and variable visual presentations. Certain lesions exhibit minimal visual contrast with the surrounding healthy tissue, creating ambiguity for the model. Similar challenges have been reported in

previous veterinary studies, where morphological variability and diverse pathological presentations reduce model performance relative to anatomical segmentation (Bortoluzzi *et al.*, 2023; Recchia *et al.*, 2025). Likewise, recent studies on bovine respiratory disease emphasize that lesion variability remains a primary limitation for artificial intelligence-based diagnostic systems (Casella *et al.*, 2023; Kamel *et al.*, 2024).

The observed variability in performance metrics likely reflects both task complexity and the limited dataset size. With only 129 images available, the model was not exposed to the full spectrum of lesion presentations encountered in field practice. This constraint increases the risk of underrepresenting specific lesion patterns, thereby reducing confidence in the model's ability to generalize to unseen cases. Consequently, while these results support the feasibility of automated lesion detection, they should be interpreted as preliminary and exploratory.

The novelty of this work lies in demonstrating that deep learning methods can identify pathological pulmonary patterns from post-mortem slaughterhouse images despite these challenging conditions. To our knowledge, few studies (Bortoluzzi *et al.*, 2023; Recchia *et al.*, 2025) have explored lesion segmentation in this specific veterinary context using practical field-acquired imagery.

Post-processing lesion quantification

An additional strength of this work lies in the post-processing algorithm developed for lesion quantification, which translates model predictions into interpretable analytical outputs. Although most computer vision studies focus primarily on classification or segmentation accuracy, quantitative analysis of lesion burden has been identified as a critical step toward practical decision-support systems in livestock production (Guarnido-Lopez *et al.*, 2024). Similarly, previous artificial intelligence frameworks for bovine respiratory disease have largely emphasized disease detection rather than quantitative lesion assessment (Casella *et al.*, 2023; Yang *et al.*, 2025).

In contrast, the proposed approach provides clinically interpretable outputs by estimating the proportions of healthy and affected lung tissue and quantifying lesion distribution separately for dorsal and ventral surfaces. This capability demonstrates the potential for standardized postmortem evaluations and supports integrating AI-assisted quantification into veterinary diagnostic workflows. Nevertheless, the accuracy of these quantitative estimates directly depends on segmentation performance: errors in boundary delineation can propagate into tissue burden calculations, particularly for small or poorly defined lesions. Therefore, validation against larger datasets and expert-annotated reference standards remains essential before these outputs can be adopted in routine veterinary decision-making.

Implications and Future Directions

The present study demonstrates the feasibility of applying deep learning techniques for anatomical segmentation, lesion detection, and lesion quantification in bovine lungs under slaughterhouse conditions. While anatomical segmentation models achieved consistent performance, lesion detection remained substantially more challenging due to lesion heterogeneity and the variability inherent to post-mortem images. These results suggest that deep learning has considerable potential for supporting veterinary post-mortem evaluations; however, the limited dataset represents a major constraint that may affect the robustness and generalizability of the reported results. Consequently, this study should be regarded as a proof-of-concept investigation, highlighting the potential of the proposed framework while underscoring the need for larger and more diverse datasets to validate its performance in real-world applications.

Recent reviews consistently identify dataset diversity, external validation, and multicenter evaluation as essential requirements for the clinical translation of deep learning models (Gupta *et al.*, 2022; Sharifani & Amini, 2023; Kufel *et al.*, 2023; Trigka & Dritsas, 2025). Accordingly, future work should prioritize the collection of substantially larger and more diverse datasets, incorporating images from multiple slaughterhouses, varied acquisition conditions, and a broader spectrum of lesion types and severities. Validation using independent external datasets will be particularly important to determine model generalizability and robustness across different operational scenarios. These efforts will be critical before such approaches can be considered for routine integration into veterinary diagnostic workflows.

CONCLUSIONS

This study establishes that deep learning enables reliable anatomical segmentation of bovine lungs from post-mortem slaughterhouse images, achieving F1-scores up to 0.827 for lung and trachea identification. However, lesion detection remains substantially more challenging (the best-performing model achieved an F1-score of 0.602), underscoring the difficulty of automated pathological pattern recognition in heterogeneous, data-limited settings. A key contribution is the development of a post-processing algorithm that converts segmentation outputs into objective lesion quantification metrics such as percentage of affected tissue (48.89% in the representative case). This framework addresses a critical gap in veterinary pathology by enabling standardized, reproducible estimation of lesion burden across different lung views. The primary limitation (a small dataset of 129 images) requires caution in interpreting performance metrics and precludes claims of generalizability. Consequently, this work should be viewed as proof-of-concept demonstrating the feasibility of AI-assisted post-mortem evaluation under real-world slaughterhouse conditions.

To translate these findings into practical tools, future work must focus on three priorities: 1) collecting large-scale, multi-site datasets, 2) conducting external validation on independent data, and 3) developing user-friendly interfaces for veterinary integration. By addressing these challenges, deep learning can become a valuable adjunct to traditional post-mortem inspection, improving disease surveillance and supporting herd health management decisions.

DECLARATIONS

Competing Interests Statement

The authors declare that they have no competing interests.

Ethics Statement

All procedures involving the animals were performed according to the guidelines established by the Bioethics Committee of the School of Veterinary Medicine and Animal Sciences of Veracruzana University (approval registration number COBIBA011/2021).

Data availability statement

The datasets and codes presented in this article are not readily available because they are part of an ongoing study. Requests to access the datasets and/or codes should be directed to Alfredo Pavón (ingfred18@gmail.com) and/or Luis Marín (luismarin@uv.mx).

Author Contributions

Conceptualization, A.P., L.M., A.L., P.G., M.B.; methodology, A.P., L.M., A.L., P.G., R.P.; software, A.P., L.M.; validation, A.P., L.M., A.L., B.D., R.P.; formal analysis, A.P., L.M., C.M., A.L., V.F.; investigation, A.P., L.M., A.L., C.M., R.P., V.F.; resources, A.L., C.M., B.D.; data curation, A.L., C.M., B.D., V.F.; writing—original draft preparation, A.P., L.M., A.L.; writing—review and editing, A.P., L.M., A.L., P.G., M.B., R.P.; visualization, A.L., V.F.; supervision, L.M., A.L., P.G., M.B., R.P., B.D.; project administration, L.M., A.L., P.G., M.B., B.D.; funding acquisition. All authors have read and agreed to the final version of the manuscript.

Funding

This research received no external funding.

REFERENCES

- Adekunle, A., & Kaniyamattam, K. (2026). Bovine respiratory disease: Epidemiological drivers, transmission dynamics, and economic implications in beef production systems. *Agriculture*, 16(3), 311. <https://doi.org/10.3390/agriculture16030311>
- Almeraya Soberanes, S., Mora Flores, J. S., García Salazar, J. A., Hernández Mendo, O., & Almeraya Quintero, S. X. (2024). Caracterización de la cadena de abasto de ganado bovino producido en la zona norte de Veracruz. *Agricultura, Sociedad y Desarrollo*, 21(2). <https://doi.org/10.22231/asdy.v21i2.1602>
- Bello, R.-W., Owolawi, P. A., Van Wyk, E. A., & Tu, C. (2024). Transfer learning-driven cattle instance segmentation using deep learning models. *Agriculture*, 14(12), 2282. <https://doi.org/10.3390/agriculture14122282>
- Blakebrough-Hall, C., McMeniman, J. P., & González, L. A. (2020). An evaluation of the economic effects of bovine respiratory disease on animal performance, carcass traits, and economic outcomes in feedlot cattle defined using four BRD diagnosis methods. *Journal of Animal Science*, 98(2), skaa005. <https://doi.org/10.1093/jas/skaa005>
- Bortoluzzi, E., Schmidt, P., Brown, R., Jensen, M., Mancke, M., Larson, R., Lancaster, P., & White, B. (2023). Image classification and automated machine learning to classify lung pathologies in deceased feedlot cattle. *Veterinary Sciences*, 10(2), 113. <https://doi.org/10.3390/vetsci10020113>
- Bumbálek, R., Zoubek, T., Ufitikirezi, J. D. D. M., Umurungi, S. N., Stehlík, R., Havelka, Z., Kunš, R., & Bartoš, P. (2025). Implementation of machine vision methods for cattle detection and activity monitoring. *Technologies*, 13(3), 116. <https://doi.org/10.3390/technologies13030116>
- Casella, E., Cantor, M. C., Setser, M. M. W., Silvestri, S., & Costa, J. H. C. (2023). A machine learning and optimization framework for the early diagnosis of bovine respiratory disease. *IEEE Access*, 11, 71164-71179. <https://doi.org/10.1109/ACCESS.2023.3291348>
- Chien, R. C., Sorensen, N. J., Payton, M. E., & Confer, A. W. (2022). Comparative histopathology of bovine acute interstitial pneumonia and bovine respiratory syncytial virus-associated interstitial pneumonia. *Journal of Comparative Pathology*, 192, 23-32. <https://doi.org/10.1016/j.jcpa.2022.01.005>
- Espinoza-Cantú, J., Rojas-Hernández, S., Olivares-Pérez, J., Romero-Rosales, T., Damian Valdez, M. A., & Aguirre-Flores, V. (2026). Evaluating perception of animal welfare among owners of family farming units on dry tropical lands. *Agro Productividad*, XII. <https://doi.org/10.32854/ewmkm125>
- Guarnido-Lopez, P., Pi, Y., Tao, J., Mendes, E. D. M., & Tedeschi, L. O. (2024). Computer vision algorithms to help decision-making in cattle production. *Animal Frontiers*, 14(6), 11-22. <https://doi.org/10.1093/af/vfae028>
- Gupta, J., Pathak, S., & Kumar, G. (2022). Deep Learning (CNN) and transfer learning: A review. *Journal of Physics. Conference Series*, 2273(1), 012029. <https://doi.org/10.1088/1742-6596/2273/1/012029>
- Houssein, E. H., Emam, M. M., & Ali, A. A. (2021). An efficient multi-level thresholding segmentation method for thermography breast cancer imaging based on improved chimp optimization algorithm. *Expert Systems with Applications*, 185, 115651. <https://doi.org/10.1016/j.eswa.2021.115651>
- Instituto Nacional de Estadística y Geografía (INEGI). (n.d.). Mexico—Censo agropecuario 2022, resultados definitivos. Retrieved April 07, 2026, from <https://www.inegi.org.mx/rnm/index.php/catalog/867>
- Jia, N., Kootstra, G., Koerkamp, P. G., Shi, Z., & Du, S. (2021). Segmentation of body parts of cows in RGB-depth images based on template matching. *Computers and Electronics in Agriculture*, 180, 105897. <https://doi.org/10.1016/j.compag.2020.105897>
- Kamel, M. S., Davidson, J. L., & Verma, M. S. (2024). Strategies for bovine respiratory disease (Brd) diagnosis and prognosis: A comprehensive overview. *Animals*, 14(4), 627. <https://doi.org/10.3390/ani14040627>
- Kufel, J., Bargiel-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning? —Examples of practical applications in medicine. *Diagnostics*, 13(15), 2582. <https://doi.org/10.3390/diagnostics13152582>
- Liu, Y., Zhang, Z., Liu, X., Wang, L., & Xia, X. (2021). Efficient image segmentation based on deep learning for mineral image classification. *Advanced Powder Technology*, 32(10), 3885-3903. <https://doi.org/10.1016/j.appt.2021.08.038>
- MSD Animal Health Republic of Ireland. (n.d.). Bovine respiratory disease. Retrieved March 28, 2026, from <https://www.msd-animal-health.ie/species/cattle/bovine-respiratory-disease/>
- O'Donoghue, S., Waters, S. M., Morris, D. W., & Earley, B. (2025). A comprehensive review: Bovine respiratory disease, current insights into epidemiology, diagnostic challenges, and vaccination. *Veterinary Sciences*, 12(8), 778. <https://doi.org/10.3390/vetsci12080778>
- Puig, A., Ruiz, M., Bassols, M., Fraile, L., & Armengol, R. (2022). Technological tools for the early detection of bovine respiratory disease in farms. *Animals*, 12(19), 2623. <https://doi.org/10.3390/ani12192623>
- Recchia, M., Scuri, L., Allegri, C., Romeo, C., Scali, F., Maisano, A. M., Santucci, G., Bontempi, G., Canesi, S., Sala, L., Recordati, C., Scanzi, E., Panseri, S., Ghidini, S., Zanardi, E., Ianieri, A., Alban, L., & Alborali, G. L. (2025). Application of computer vision for automated detection of different lesions in pig lungs: An exploratory study. *Preventive Veterinary Medicine*, 245, 106672. <https://doi.org/10.1016/j.prevetmed.2025.106672>
- Rossi, C. A. S., Grossi, S., Fortuna, M., Schiavon, E., Fava, E., Adami, S., & Compiani, R. (2022). Sanitary, environmental and nutritional management to reduce the incidence of bovine respiratory disease and the use of antibiotics in fattening beef cattle. *Large Animal Review*, 28(2), 65-72. <https://www.largeanimalreview.com/index.php/lar/article/view/451>
- Sharifani, K., & Amini, M. (2023). Machine learning and deep learning:

- A review of methods and applications. *World Information Technology and Engineering Journal*, 10(7), 3897–3904. <https://ssrn.com/abstract=4458723>
- Singh, D., Singh, R., Gehlot, A., Bhandari, G., Verma, A., Gangwar, P., & Pandey, P. S. (2025). Sensor and computer vision based cattle health monitoring and management. *International Journal of Electrical and Electronics Engineering*, 12(1), 94–103. <https://doi.org/10.14445/23488379/IJEEE-V12I1P109>
- Trigka, M., & Dritsas, E. (2025). A comprehensive survey of machine learning techniques and models for object detection. *Sensors*, 25(1), 214. <https://doi.org/10.3390/s25010214>
- Ultralytics. (n.d.-a). Data Augmentation Using YOLO Ultralytics. Retrieved April 10, 2026, from <https://docs.ultralytics.com/guides/yolo-data-augmentation/>
- Ultralytics. (n.d.-b). Performance metrics deep dive. Retrieved April 11, 2026, from <https://docs.ultralytics.com/guides/yolo-performance-metrics/TABLES>
- Wu, D., Han, M., Song, H., Song, L., & Duan, Y. (2023). Monitoring the respiratory behavior of multiple cows based on computer vision and deep learning. *Journal of Dairy Science*, 106(4), 2963–2979. <https://doi.org/10.3168/jds.2022-22501>
- Xiao, S., Dhand, N. K., Wang, Z., Hu, K., Thomson, P. C., House, J. K., & Khatkar, M. S. (2025). Review of applications of deep learning in veterinary diagnostics and animal health. *Frontiers in Veterinary Science*, 12, 1511522. <https://doi.org/10.3389/fvets.2025.1511522>
- Yang, S. X., Han, Y., Ma, W., Tulpan, D., Li, J., Li, J., & Yue, Y. (2025). Review of computer vision for livestock body conformation assessment. *Agriculture Communications*, 3(3), 100099. <https://doi.org/10.1016/j.agr-com.2025.100099>