Do AI and Human Experts See Pneumonia the Same Way?

Davide Beltrame [*1] Giacomo Cirò [*1]

[1]Bocconi Students for Machine Learning, Bocconi University, Milan, Italy
{davide.beltrame giacomo.ciro}@studbocconi.it

June 16, 2025

Abstract

We examine the alignment between deep learning models and human expert reasoning for pneumonia diagnosis in pediatric chest X-rays. We train four CNN architectures (AlexNet, VGG-16, ResNet-50, InceptionNet-V1) on 5,216 images and compare their saliency maps against consensus annotations from 14 medical experts using Intersection over Union (IoU) and Pointing Game metrics. While pretrained models consistently outperform non-pretrained variants, with ResNet-50 achieving the best accuracy (94.87%), diagnostic performance does not correlate with explainability. VGG-16 produces the most expert-aligned saliency maps (22.19% IoU, on par with inter-expert agreement), while ResNet-50's superior accuracy corresponds to worse interpretability. Our findings demonstrate that high classification accuracy alone is insufficient for clinical interpretability, highlighting the need for evaluation frameworks that jointly consider both predictive performance and human-interpretable explanations in medical AI. Our code is available on GitHub[1].

1 Introduction

Deep learning has achieved state-of-the-art performance in medical image analysis, particularly in tasks such as disease classification [9, 10]. However, the opacity of its decision-making remains a barrier to clinical adoption: medical professionals demand transparency to verify that predictions are grounded in medically relevant evidence [3].

This study assesses the alignment of four CNN architectures—AlexNet [5], VGG-16 [15], ResNet-50 [2], and InceptionNet-V1 [17]—with human experts for pneumonia classification, using saliency-based explanation techniques [11].

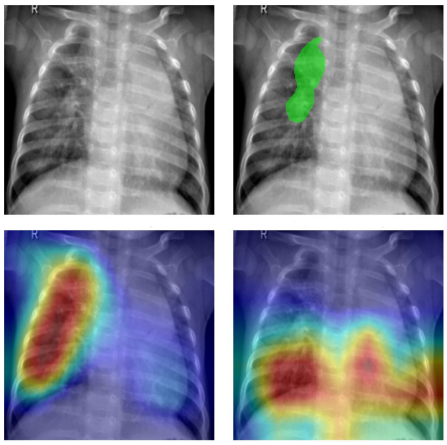


Figure 1: Saliency Map Example. Top row shows the original X-ray and a raw expert annotation. Bottom row displays CAM for VGG-16 (left) and ResNet-50 (right).

We train few variants of each architecture to perform binary classification of X-ray images. We explore training from scratch or fine-tuning from ImageNet [1] checkpoints, as well as adapting the classification head with a Global Average Pooling layer [6] or not.

We generate visual explanations of the best models' predictions using Class Activation Mapping (CAM) [19] and Gradient-weighted CAM (Grad-CAM) [13]. Then, we ask a range of medical experts to annotate diagnostically relevant regions. We aggregate these annotations into one *consensus map* per input image, and compare with the models' maps using Intersection over Union

[*]Equal contribution, the ordering is alphabetical.
[1]https://github.com/davide-beltrame/medimg-saliency-benchmark

(IoU) and Pointing Game (PG) metrics.

In terms of diagnostic accuracy, we find that pretrained models consistently outperform their non-pretrained counterparts across all architectures. In terms of explainability, we observe that the most accurate classifier does not align best with expert reasoning, and some high-performing models exhibit poor interpretability.

Our findings show that diagnostic accuracy alone is an insufficient proxy for explainability and highlight the need for developing evaluation frameworks that jointly consider both predictive performance and interpretability. Achieving trustworthy AI in medical imaging requires explicit optimization for human-interpretable explanations, not merely high classification accuracy.

Table 1: Annotations. Statistics and expert agreement metrics.

Test Images	50
Annotations (Valid)	318 (281)
Annotations per Image	5.62
Unique Annotators	14
Non-trivial Consensus Maps	50
Active Pixels per Annotation (avg.)	9.4 %
Expert-Expert IoU	23.32%
Expert-Random IoU	3.29%
P-value	< 0.001

2 Problem Analysis

We source pediatric chest X-ray images–showing signs of interstitial pneumonia, often diffuse and non-localized–from a previous study investigating image-based deep learning to identify medical diagnoses [4]. In particular, we obtain 5,856 images labeled as either pneumonia or normal. We follow a similar data split as the original work: 3,875 pneumonia and 1,341 normal images for training; 390 pneumonia and 234 normal images for testing. The original validation set contains only 16 images, which we consider insufficient, thus we reallocate 10% of the training set for validation. We simplify the task to binary classification, disregarding the original distinction between bacterial and viral pneumonia.

The main challenges include class imbalance, limited dataset size, and potential dataset-specific artifacts. We hypothesize that CNNs can achieve high accuracy and that fine-tuned models will outperform scratch-trained ones, with saliency maps meaningfully aligning with expert annotations.

3 Method

3.1 Model Development

We implement four CNN architectures: AlexNet (baseline), VGG-16 (deeper traditional architecture), ResNet-50 and InceptionNet-V1 (advanced models). For each architecture, we evaluate a version trained from scratch and one initialized from ImageNet pretraining [1].

To reduce overfitting, we apply data augmentation by randomly rotating, scaling, translating, and color-jittering each input image. Additionally, with 50% probability, we apply edge sharpening using the transformation $2 \cdot I - G$, where I is the original image and G is a blurred version obtained with a 3×3 uniform kernel. The hyperparameters used for augmentation (see Section 4.1) are either empirically tuned or adopted from PyTorch default settings [8]. We do not apply normalization, as it consistently degrades performance across all models.

To address the dataset imbalance (approximately 75% pneumonia class), we implement weighted random sampling and ensure each mini-batch contains an equal proportion of both classes.

For saliency generation, we use CAM [19] (exploiting the final convolutional layer and classification weights), and Grad-CAM [13] (using gradients flowing into the final convolutional layer). We normalize all generated maps to [0,1] and binarize them using a fixed threshold of 0.5. All maps are up-sampled to 224×224 pixels for fair comparison.

While Grad-CAM is architecture-agnostic and compatible with any CNN, CAM requires the model to end with a Global Average Pooling (GAP) layer [6] followed by a linear classifier. This structure is already present in ResNet-50 and InceptionNet-V1, while AlexNet and VGG-16 must be adapted accordingly. We evaluate both the original and the modified versions of these architectures.

3.2 Expert Alignment

To investigate model-expert alignment, we compare the saliency maps produced by the trained models with a *consensus map* derived from expert annotations (see Table 1).

We invite medical experts to annotate a randomly sampled subset of 50 pneumonia-positive images from the test set, marking regions they considered relevant for diagnosis via a web-based interface (see Figure 9, 10, 11).

We assess annotation quality by measuring inter-expert agreement using pairwise IoU scores,

Table 2: Chosen Models Performance. Confidence intervals at 95% are estimated on the test set with 1000 bootstrap samples. We highlight in bold the best results overall.

Model	Adapted	Params.	Pretr.	Acc.	F1	AU-ROC	Spec.
AlexNet	Yes	2.5M	Yes	90.34 ± 2.24	92.72 ± 1.79	96.72 ± 1.13	76.40 ± 5.06
VGG-16	Yes	14.7M	Yes	91.46 ± 2.24	93.50 ± 1.78	97.93 ± 0.96	79.85 ± 5.03
InceptionNet-V1	-	12M	Yes	91.84 ± 2.24	93.79 ± 1.76	98.06 ± 0.97	80.37 ± 5.19
ResNet-50	-	23.5M	Yes	**94.87** ± 1.68	**96.00** ± 1.39	**98.24** ± 1.00	**88.86** ± 3.86

and comparing the average to a baseline derived from expert–random annotation pairs. To generate random annotations, we create binary grids matching expert annotation density and upsample them (see Figure 2). We use a Mann-Whitney U test on the distribution of pairwise IoU scores to test the null hypothesis that the expert-expert agreement scores come from the same distribution of the expert-random ones.

To construct the *consensus map* for each test image, we compute the pixel-wise average of the collected annotations and binarize this averaged map using a fixed threshold.

Model–expert alignment is evaluated using the Intersection over Union (IoU) and Pointing Game (PG) metrics. IoU measures the overlap between the binarized saliency map and the *consensus map*, while PG checks whether the most activated pixel in the saliency map falls within the consensus region. To limit redundancy, we evaluate these metrics only on one model variant per architecture.

4 Experiments

4.1 Model Development

We train all models for 10 epochs (batch size 64) with early stopping after 3 epochs without validation improvement. We use AdamW optimization [7] ($\beta_2 = 0.95$ [18]), gradient clipping (norm 0.5)[14], and one-cycle learning rate scheduling [16] (max 10^{-4}, 30% warm-up).

The parameters for data augmentation are reported in Table 5. All images are downsampled to 224×224 pixels and normalized to $[0, 1]$ before processing.

4.2 Expert Alignment

Our 14 annotators range from medical students to experienced radiologists (see Appendix A, Figure 3). On average, each image received 5.62 annotations (see Appendix A, Figure 4).

As shown in Table 1, inter-expert agreement is significantly above chance, validating the overall quality of the annotations (see also Appendix A, Figure 5).

We generate *consensus maps* by averaging the annotations and applying a fixed binarization threshold of 0.5. For every test image, at least 50% of annotators marked overlapping regions.

5 Results

Our experiments show that pretrained models consistently outperform their non-pretrained counterparts across all architectures (see Appendix A Table 4). This highlights the effectiveness of transfer learning: features learned from natural images transfer well to medical imaging tasks.

ResNet-50 with pretraining achieves the best overall performance. However, without pretraining, its performance drops sharply, underscoring the importance of proper initialization for deeper architectures when training data is limited. VGG-16 and InceptionNet-V1 also benefit substantially from pretraining, while AlexNet shows relatively stable performance even when trained from scratch, suggesting that simpler models may generalize better with limited data.

For consistency, we select the models implemented with the GAP layer followed by the linear classifier (see Table 2). The parameter savings in GAP-based versions are substantial and the performance trade-off is marginal.

In terms of explainability (see Table 3), VGG-16 consistently achieves the highest IoU scores, indicating that its saliency maps most closely align with expert annotations—approaching the level of inter-expert agreement.

However, in terms of PG, which captures the model's ability to localize the most diagnostically relevant point, InceptionNet-V1 and VGG-16 lead under CAM, while InceptionNet-V1 scores highest under Grad-CAM. This suggests that

Table 3: Model-Expert Agreement. Entries show IoU and PG between model-generated saliency maps and expert annotations, along with p-values. The threshold for map binarization is fixed at 0.5.

Model	GradCAM		CAM		Random	
	IoU	PG	IoU	PG	IoU	PG
AlexNet	9.08 0.000	28.0 0.000	6.78 0.756	12.0 0.006	5.93 -	0.0 -
InceptionNet-V1	16.33 0.000	**38.0** 0.000	16.33 0.000	38.0 0.000	5.36 -	0.0 -
ResNet-50	11.46 0.000	20.0 0.001	10.84 0.000	20.0 0.001	4.99 -	0.0 -
VGG-16	**19.6** 0.000	32.0 0.000	**22.19** 0.000	**40.0** 0.000	4.96 -	0.0 -

InceptionNet-V1 is particularly effective at pinpointing critical regions.

In contrast, AlexNet performs poorly across both metrics and saliency methods, with its IoU under CAM nearly equivalent to random performance.

6 Discussion

High diagnostic accuracy does not guarantee interpretability: the best performing model (Resnet-50) is not the most aligned, and a model with reasonable classification accuracy (AlexNet) showed poor alignment.

We also observe a meaningful divergence between IoU and PG scores for more aligned models, likely driven by architectural differences. These patterns, exemplified in Figure 1, emphasize the need for multi-dimensional evaluation frameworks that consider both accuracy and alignment with expert reasoning.

While no standard benchmark exists for our task, our IoU scores are competitive with prior work [12].

7 Conclusions

This research contributes to explainable AI in medical imaging by systematically assessing how CNN-based pneumonia detection models align with expert reasoning.

We find that strong diagnostic performance does not guarantee interpretability, suggesting the need to develop methods that explicitly align decision-making with clinical expertise.

A preliminary analysis suggests that physicians produce more consistent annotations than students, underscoring the need to examine the role of expertise in establishing reliable ground truth. As our study focuses on a homogeneous dataset of pediatric chest X-rays, future work should assess the generalizability of these findings across diverse patient demographics, anatomical regions, and pathological conditions.

Our threshold sensitivity analysis (see Appendix A, Figures 6 and 7) highlights the need to explore different thresholding methods in explainability evaluations.

We recommend: (1) collecting more annotations and exploring different aggregation strategies; (2) expanding the range of models to measure the correlation between accuracy and alignment; (3) testing additional saliency methods and binarization techniques; and (4) incorporating dual-view radiographs (frontal and lateral).

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[4] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[6] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

[9] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[11] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[12] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest x-ray interpretation. *medRxiv*, 2021.

[13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[14] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[16] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[18] Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36:10271–10298, 2023.

[19] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

A Appendix

Table 4: Models Performance. Confidence intervals at 95% are estimated on the test set with 1000 bootstrap samples. We highlight in bold the best results per model, and in bold plus underline the best results overall.

Model	Adapted	# Param.	Pretrained	Accuracy	F1	ROC AUC	Specificity
AlexNet	**No**	57M	No	87.42 ± 2.56	90.29 ± 2.11	94.50 ± 1.72	76.84 ± 5.34
			Yes	**91.98** ± 2.16	**93.75** ± 1.72	**97.01** ± 1.08	**84.60** ± 4.55
	Yes	2.5M	No	87.48 ± 2.48	90.22 ± 2.05	93.40 ± 1.88	78.98 ± 5.11
			Yes	90.34 ± 2.24	92.72 ± 1.79	96.72 ± 1.13	76.40 ± 5.06
VGG-16	**No**	134M	No	82.64 ± 2.88	87.53 ± 2.27	93.73 ± 1.89	57.65 ± 6.13
			Yes	**92.12** ± 2.09	**93.87** ± 1.72	97.72 ± 0.95	**84.56** ± 4.34
	Yes	14.7M	No	86.35 ± 2.73	89.05 ± 2.33	92.46 ± 2.04	82.01 ± 4.81
			Yes	91.46 ± 2.24	93.50 ± 1.78	**97.93** ± 0.96	79.85 ± 5.03
InceptionNet-V1	-	12M	No	87.92 ± 2.49	91.11 ± 1.94	96.56 ± 1.30	69.14 ± 5.57
			Yes	**91.84** ± 2.24	**93.79** ± 1.76	**98.06** ± 0.97	**80.37** ± 5.19
ResNet-50	-	23.5M	No	62.77 ± 3.69	77.02 ± 2.76	82.82 ± 3.23	0.82 ± 1.06
			Yes	**94.87** ± 1.68	**96.00** ± 1.39	**98.24** ± 1.00	**88.86** ± 3.86

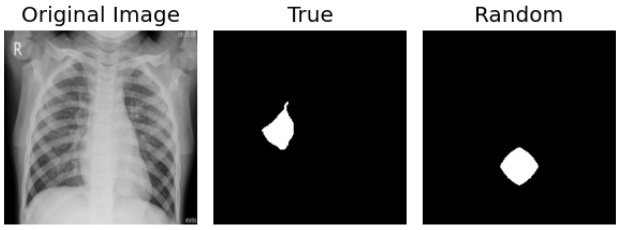


Figure 2: True vs Random Annotation. Selected example to show true and randomly generated annotation. We estimate the average number of active pixels in valid expert masks, create a small binary grid with the same density of active pixels, upsample it via bilinear interpolation, and binarize the result

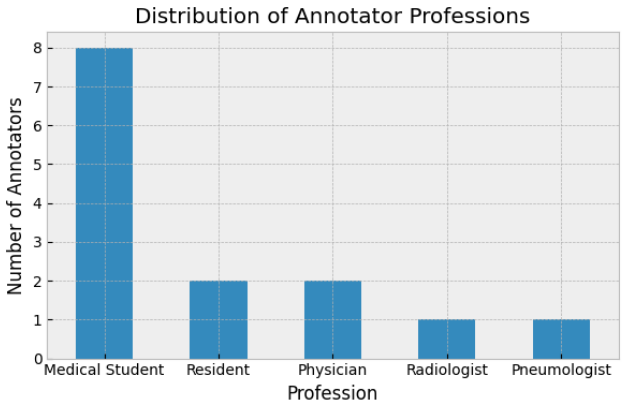


Figure 3: Professions. The distribution of our annotators' professions.

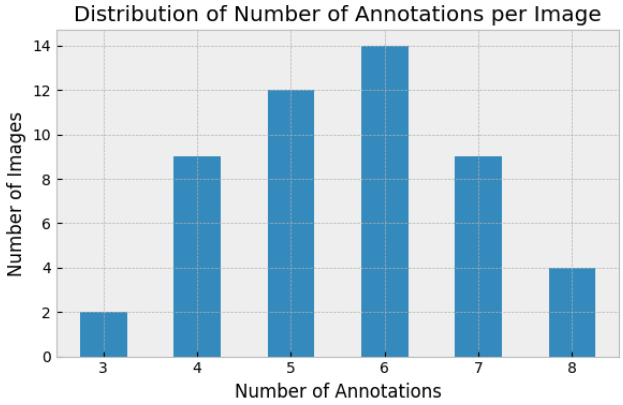


Figure 4: Annotations Distribution. The number of annotations collected for each image.

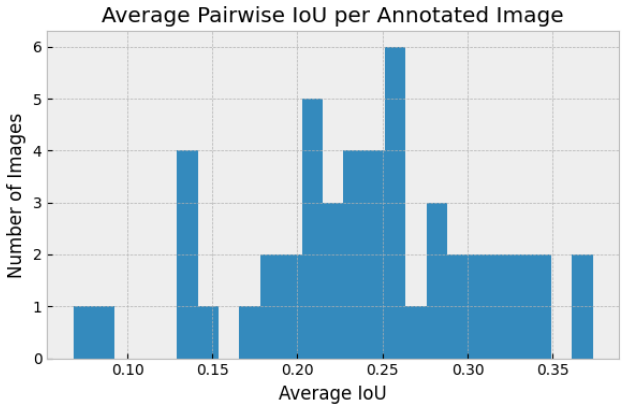


Figure 5: Average IoU. For each image, we measure the IoU between all pairs of annotations for that image and then compute the average.

Augmentation Type	Parameter Range
Rotation	$\alpha \in [0, 15°]$
Translation	$\mathbf{t} \in [-22.4, 22.4]^2$
Scaling	$\beta \in [0.9, 1.1]$
Brightness	$\gamma \in [0.7, 1.3]$
Contrast	$\delta \in [0.7, 1.3]$
Saturation	$\epsilon \in [0.7, 1.3]$
Hue	$\zeta \in [0.9, 1.1]$

Table 5: Data augmentation strategy and parameter ranges.

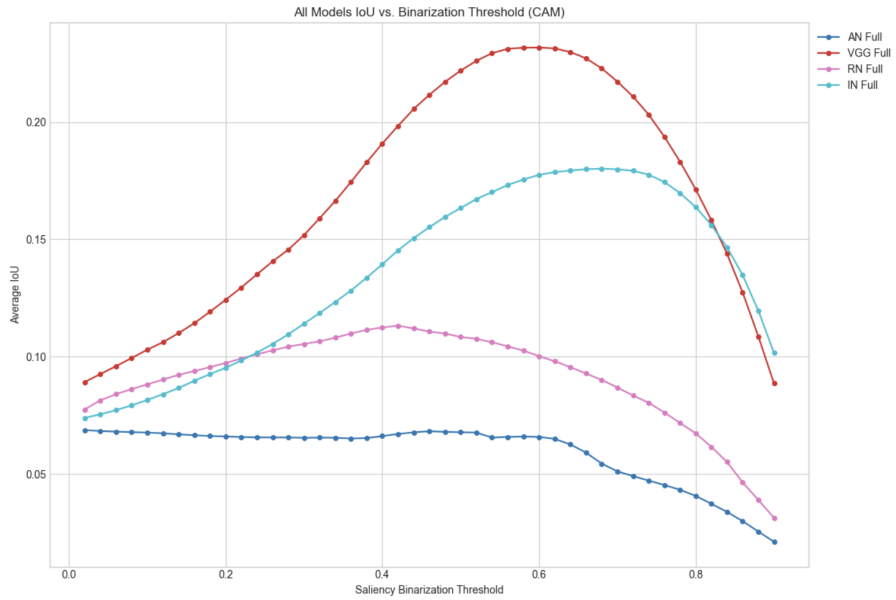


Figure 6: Model–Expert IoUs vs. Thresholds (CAM). IoU between model saliency maps and expert annotations as a function of the binarization threshold applied to CAM outputs.

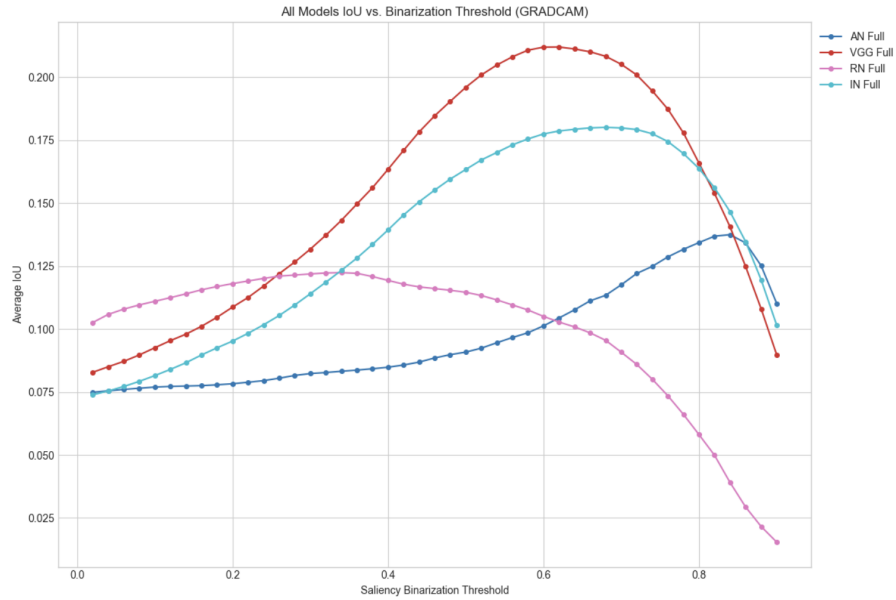


Figure 7: Model–Expert IoUs vs. Thresholds (Grad-CAM). IoU between model saliency maps and expert annotations as a function of the binarization threshold applied to Grad-CAM outputs.

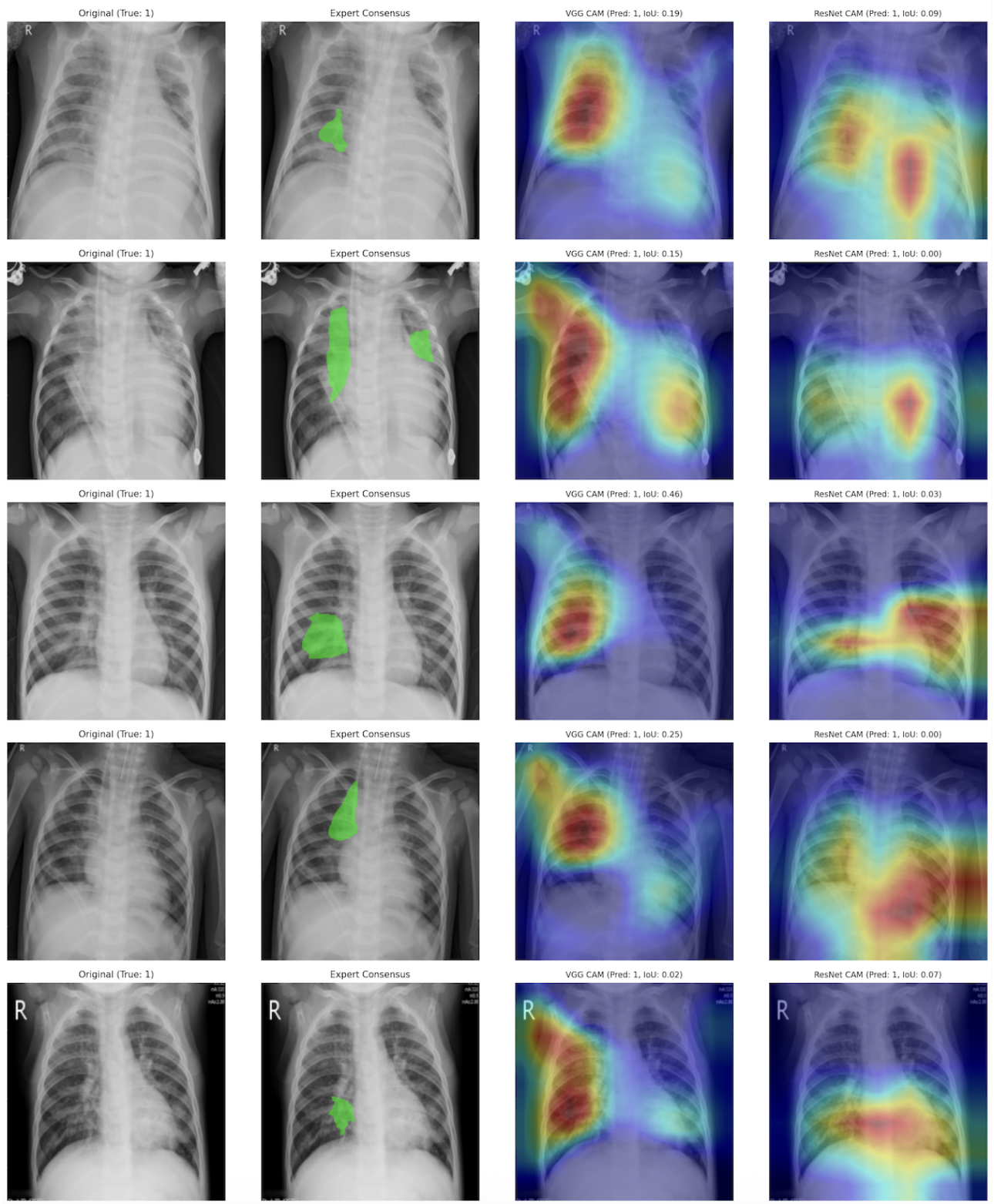
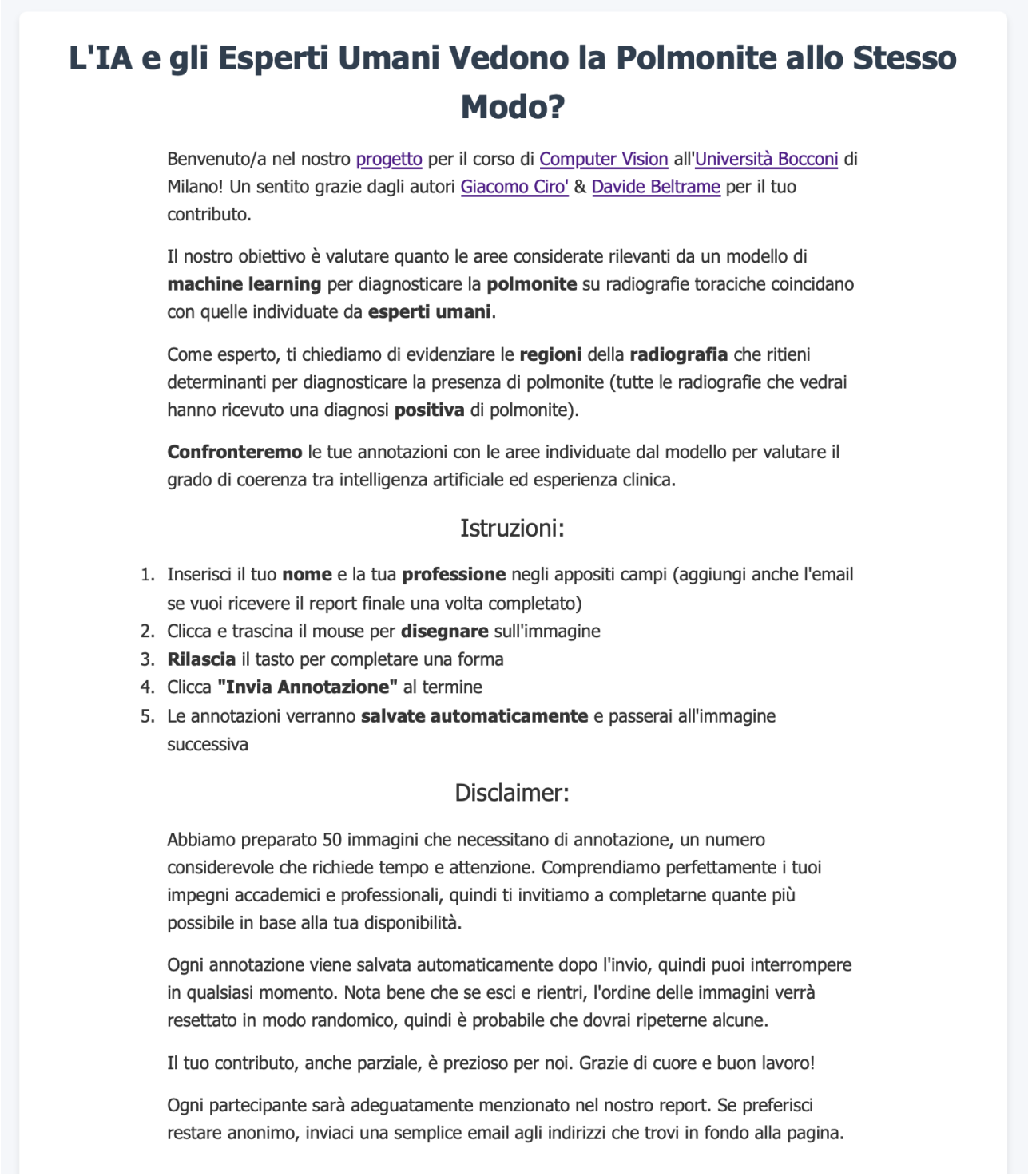


Figure 8: Saliency Map Comparison. Visual comparison between expert annotations and model-generated saliency maps for five pneumonia cases. For each case, from the left: original X-ray, expert consensus annotation (green overlay), VGG-16 CAM, and ResNet-50 CAM.



L'IA e gli Esperti Umani Vedono la Polmonite allo Stesso Modo?

Benvenuto/a nel nostro progetto per il corso di Computer Vision all'Università Bocconi di Milano! Un sentito grazie dagli autori Giacomo Ciro' & Davide Beltrame per il tuo contributo.

Il nostro obiettivo è valutare quanto le aree considerate rilevanti da un modello di **machine learning** per diagnosticare la **polmonite** su radiografie toraciche coincidano con quelle individuate da **esperti umani**.

Come esperto, ti chiediamo di evidenziare le **regioni** della **radiografia** che ritieni determinanti per diagnosticare la presenza di polmonite (tutte le radiografie che vedrai hanno ricevuto una diagnosi **positiva** di polmonite).

Confronteremo le tue annotazioni con le aree individuate dal modello per valutare il grado di coerenza tra intelligenza artificiale ed esperienza clinica.

Istruzioni:

1. Inserisci il tuo **nome** e la tua **professione** negli appositi campi (aggiungi anche l'email se vuoi ricevere il report finale una volta completato)
2. Clicca e trascina il mouse per **disegnare** sull'immagine
3. **Rilascia** il tasto per completare una forma
4. Clicca **"Invia Annotazione"** al termine
5. Le annotazioni verranno **salvate automaticamente** e passerai all'immagine successiva

Disclaimer:

Abbiamo preparato 50 immagini che necessitano di annotazione, un numero considerevole che richiede tempo e attenzione. Comprendiamo perfettamente i tuoi impegni accademici e professionali, quindi ti invitiamo a completarne quante più possibile in base alla tua disponibilità.

Ogni annotazione viene salvata automaticamente dopo l'invio, quindi puoi interrompere in qualsiasi momento. Nota bene che se esci e rientri, l'ordine delle immagini verrà resettato in modo randomico, quindi è probabile che dovrai ripeterne alcune.

Il tuo contributo, anche parziale, è prezioso per noi. Grazie di cuore e buon lavoro!

Ogni partecipante sarà adeguatamente menzionato nel nostro report. Se preferisci restare anonimo, inviaci una semplice email agli indirizzi che trovi in fondo alla pagina.

Figure 9: Annotation Platform. [1/3]

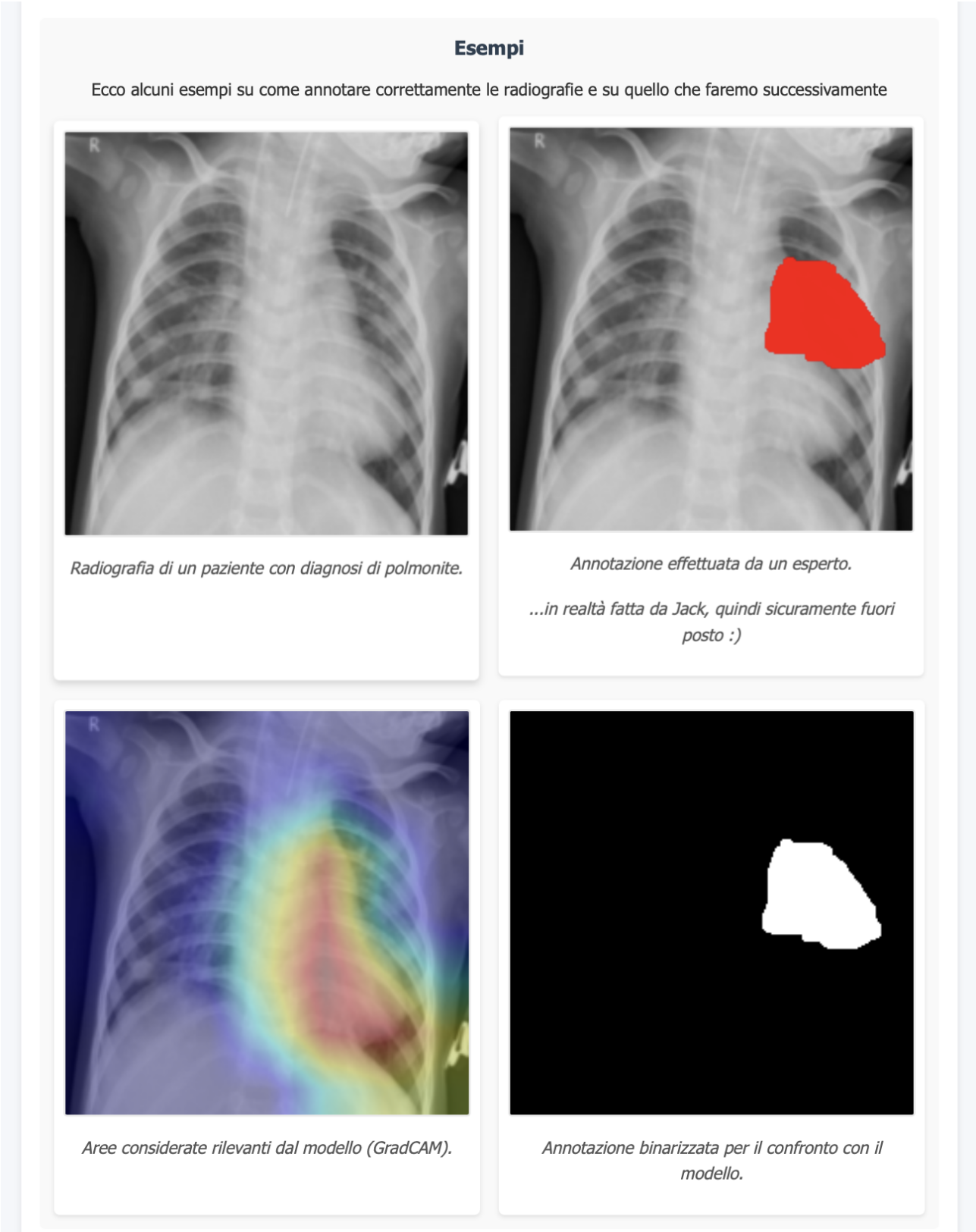


Figure 10: Annotation Platform. [2/3]

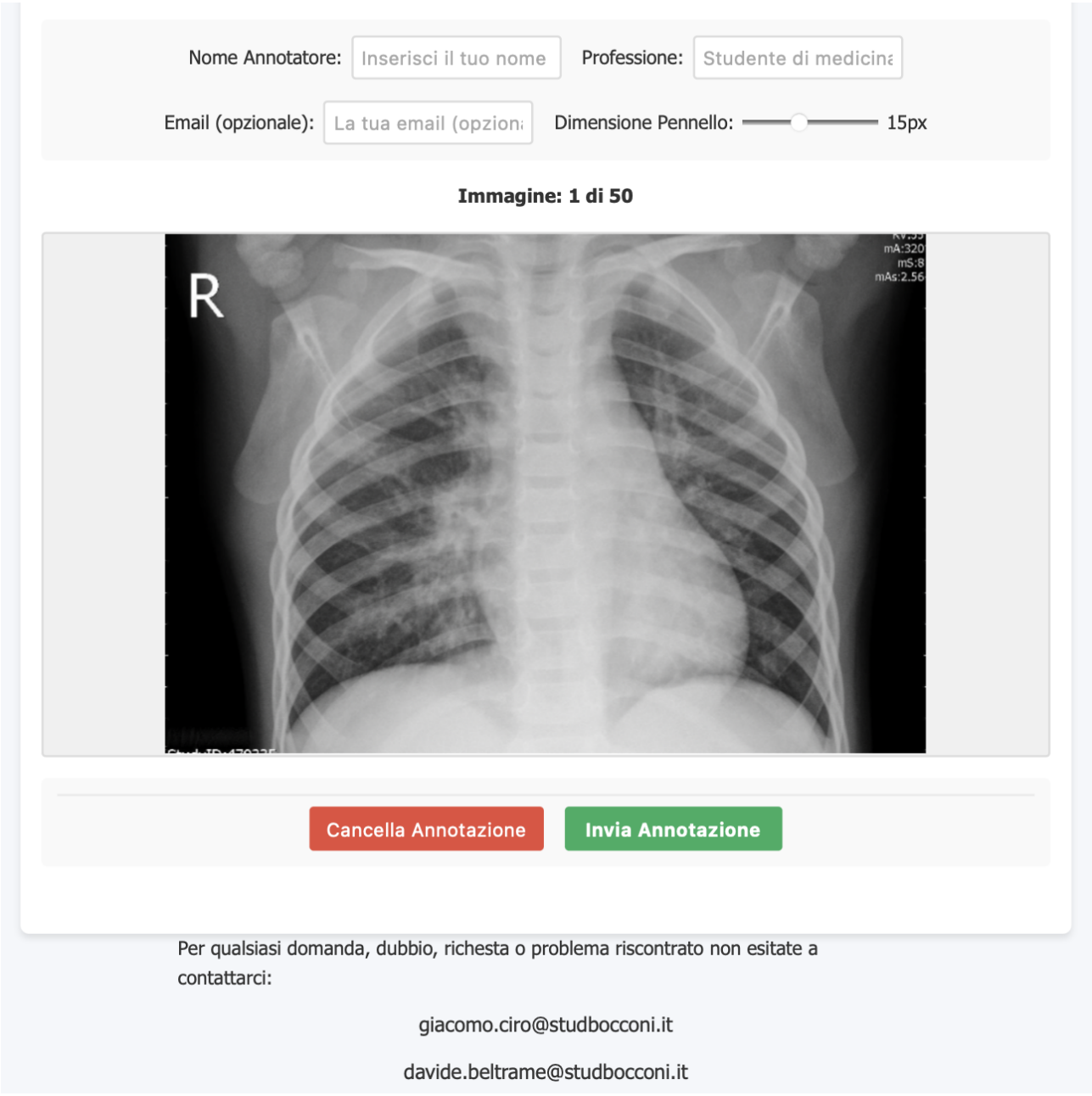


Figure 11: Annotation Platform. [3/3]