ROMANIAN ACADEMY

School of Advanced Studies of the Romanian Academy

„Simion Stoilow" Institute of Mathematics

PHD THESIS

Deep visual learning models for medical imaging and analysis

THESIS ADVISOR:

Prof. Dr. Marius Leordeanu

PHD STUDENT:

Florin Condrea

2025

Contents

ABSTRACT

Weaving Anatomical Knowledge into Computer Vision for Data-Limited Medical Applications

"It's still magic even if you know how it's done." – Terry Pratchett

Medical artificial intelligence has emerged as a transformative field at the intersection of computer vision, healthcare, and machine learning, enabling improved diagnostic accuracy and efficiency in clinical settings. As medical imaging technologies continue to advance, the ability to accurately analyze and interpret complex radiological data has become increasingly important for enhancing patient outcomes, reducing diagnostic errors, and optimizing clinical workflows. This thesis advances the field of medical AI through a series of innovative deep learning approaches focusing on pulmonary embolism detection and vital sign monitoring with limited human supervision.

We begin by approaching the medical task of pulmonary embolism detection with a novel anatomically-aware dual-hop framework for pulmonary embolism detection that leverages organ segmentation to improve detection accuracy. This approach significantly improves upon existing methods by integrating anatomical context into the detection pipeline, achieving state-of-the-art results on a large scale public pulmnoary embolism detection dataset. To address the critical challenge of annotation granularity, we introduce Label Up, a weakly supervised learning pipeline that transforms image-level annotations into detailed segmentation masks through model explainability techniques, effectively bridging the gap between coarse annotations and fine-grained segmentations without requiring additional manual annotation effort.

The thesis further addresses the challenge of data scarcity through an innovative synthetic data generation approach specifically designed for medical imaging applications. Our method preserves the temporal statistical properties of real-world medical data while main-

taining privacy constraints, while requiring no human labeling. For vital sign monitoring, we develop deep learning models capable of accurate and continuous assessment of patient physiological parameters from non-invasive sensors.

Our research advances the field through technical innovations that address key challenges in medical AI deployment: anatomical information integration for improved contextual understanding, weakly supervised learning pipelines to maximize the utility of limited annotations, and synthetic data generation to overcome data scarcity. These contributions collectively enhance the performance, interpretability, and clinical applicability of deep learning models in medical imaging analysis, paving the way toward more effective computer-aided diagnostic systems that can meaningfully support clinical decision-making and improve patient care.

Keywords – *Deep Learning, Healthcare AI, Computer-Aided Detection, Artificial Intelligence, Anatomically-Aware Models, Weakly Supervised Learning, Synthetic Data Generation, Explainable AI, Medical Imaging, Pulmonary Embolism, Vital Sign Monitoring, Clinical Decision Support*

Chapter 1

INTRODUCTION

1.1 Deep learning in the medical field

Artificial intelligence has transformed healthcare by enabling complex medical data analysis. The evolution from early expert systems to modern deep learning applications reflects broader AI advancements driven by computational power, algorithmic innovation, and data availability[36]. Healthcare presents unique AI implementation challenges due to complex, heterogeneous, privacy-sensitive data requiring specialized approaches[38], alongside regulatory requirements and integration challenges with established clinical practices[14].

Recent medical AI developments have accelerated through improved computing infrastructure, electronic health records implementation, and machine learning advances[11]. These foundations have enhanced precision in medical imaging analysis and diagnostic support[21], with the field moving toward AI-powered decision support tools and personalized medicine approaches that address healthcare delivery gaps while improving patient outcomes[10].

Deep learning has particularly impacted medical applications, with convolutional neural networks revolutionizing image analysis and recurrent architectures enabling sophisticated clinical text and time-series data processing[28]. However, clinical deployment remains limited due to interpretability, data quality, generalizability, and regulatory compliance challenges[18].

The healthcare landscape continues to evolve with AI solutions strengthening professional-patient relationships, moving from telemedicine toward integrated intelligent systems enhancing clinical decision-making while preserving human care[36].

1.2 Motivation

Medical AI research represents a unique intersection of ethical responsibility and technical innovation. Healthcare AI can transform patient outcomes, enhance clinical decision-making, and address global healthcare disparities[36]. From an ethical perspective, healthcare is ideal for AI research, offering tangible benefits through early disease detection, improved diagnostic accuracy, and personalized treatment plans[14].

Healthcare AI addresses accessibility and equity challenges globally. In regions with limited specialist access, AI systems can provide otherwise unavailable diagnostic support and triage capabilities[10]. This ethical dimension provides both a moral compass and purpose transcending purely technical achievements.

Technically, medical AI presents compelling research challenges involving complex, high-dimensional multimodal data requiring sophisticated modeling approaches[11]. Medical applications demand exceptional accuracy, reliability, and interpretability, pushing machine learning boundaries and necessitating innovations in model design and validation protocols[38].

Medical AI research benefits from clearly defined tasks with objective evaluation criteria, whether detecting pulmonary embolisms, predicting disease progression, or generating synthetic training examples for rare conditions[21]. This clarity enables focused research with direct clinical translation pathways.

1.3 Challenges

Several key challenges define the current medical AI research landscape:

Data Scarcity: Annotated medical datasets are limited due to high labeling costs and expertise requirements, particularly problematic for rare diseases[16]. For conditions with prevalence below 1 in 2,000 individuals, insufficient imaging data leads to poor model generalization[11].

Privacy Concerns: Stringent privacy regulations restrict patient data sharing, necessitating federated learning or synthetic data generation techniques[30]. GDPR and HIPAA impose strict requirements creating multi-institutional collaboration barriers[26], catalyzing privacy-preserving techniques including differential privacy and homomorphic encryption[17].

Interpretability: Deep learning models' black-box nature hinders clinical adoption where

transparency is crucial[15]. Clinicians require not only accurate predictions but explanations aligning with medical reasoning[38]. Approaches like attention mechanisms and concept-based explanations aim to bridge algorithmic complexity and clinical interpretability[35].

Bias and Fairness: Imbalanced datasets lead to biased predictions affecting certain demographic groups disproportionately[12]. Historical biases in medical practice and healthcare access encoded in training data potentially perpetuate existing disparities[4], with models trained on specific populations performing poorly on underrepresented groups[20].

1.4 Our main contributions

This thesis makes significant contributions to medical artificial intelligence, addressing healthcare application challenges while advancing methodologies for robust, clinically relevant AI systems:

End-to-End Diagnostic Systems: We explore end-to-end diagnostic medical AI pipelines leveraging medical information, using anatomical information for data preprocessing and model pretraining in Chapter 2, and clinical information for synthetic data generation in Chapter 5.

Tackling data scarcity: We developed two frameworks enabling machine learning model training on complex tasks without requiring human annotations:

Our Label Up framework[8] upgrades annotation granularity by leveraging model explainability, applied to pulmonary embolism detection by upgrading image-level annotations to voxel-level annotations enabling individual emboli localization.

In Chapter 5, we present synthetic data generation incorporating medical priors for remote vital sign monitoring. Our synthetic thermal videos use simple geometric primitives transformed based on physiologically accurate frequencies. Models trained exclusively on this synthetic data demonstrate remarkable generalization to real thermal videos, validating that effective training requires only essential temporal concepts isolated from complex visual features.

Architectural improvements: We introduce an architecture leveraging both global and local information in our HopNet pipeline (Chapter 2), utilizing multiple data passes while aggregating features from previous passes as input for subsequent passes.

Clinical validation: We evaluate our work in clinical settings (Chapter 3), examining

actual medical value for patients with detailed clinical perspective results demonstrating real-world human impact.

Chapter 2

ANATOMICALLY INFORMED MODELS

Pulmonary embolism (PE) remains a leading cause of cardiovascular mortality, with prompt and accurate diagnosis being crucial for patient outcomes. While computed tomographic pulmonary angiography (CTPA) serves as the gold standard for PE diagnosis, the complexity and subtlety of emboli manifestations pose significant challenges for radiologists. To address these challenges, In this chapter we introduce HopNet, a novel 3 stage deep learning approach that leverages anatomical awareness and a dual-hop architecture for improved PE detection in CTPA scans, obtaining state of the art performance on a large scale public dataset [6] for PE slice level classification. Our solutions are built upon a strong baseline model based on EfficientNetV2 [34], and our contributions are applied along three independent axes of development, which prove to be necessary for an accurate performance. An overview of our 3 stage pipeline is displayed in Figure 2.1.

First stage: anatomically aware masking and cropping of lung and heart regions. Deep neural modules trained on physiological information for segmenting lung and heart regions are used to segment only the relevant information with respect to PE detection. By leveraging cropped volumes, we are able to upsample the resolution of the relevant anatomical structures while keeping the original input size. An added benefit is the removal of external structures that may be mistaken for PEs. The preprocessing brings an 0.6 % F1 score improvement on

This section is based on the paper: Condrea, Florin, Marius Leordeanu et al. "Anatomically aware dual-hop learning for pulmonary embolism detection in CT pulmonary angiograms." Computers in Biology and Medicine 174 (2024): 108464.

study level over using non-cropped volumes at the same runtime.

Second stage: anatomically aware pretraining. Relevant features are pretrained on the task of localizing specific anatomical landmarks, before starting the PE learning stage. This stage has the role of priming the neural network with both anatomical knowledge, aiding in differentiating the various complex anatomical structures, as well as learning the particularities of medical imaging. Brings an 0.4 % F1 score improvement on study level over the strong Imagenet21K [29] pretraining.

Third stage: dual-hop architecture for PE detection. The dual-hop architecture performs classification in two-phases. The first phase performs an initial evaluation, and the second phase, having access to the initial input as well as the output of the first phase, is able to produce a more accurate, refined prediction. This architectural improvement has the key benefit of providing global highly refined features at the beginning of the ulterior phases. This approach simulates a very large receptive field early on, which combined with the refined features, obtains an 1.6 % F1 score on study level improvement over single hop architecture.

Our final model obtains state-of-the-art performance, obtaining an F1 score on study

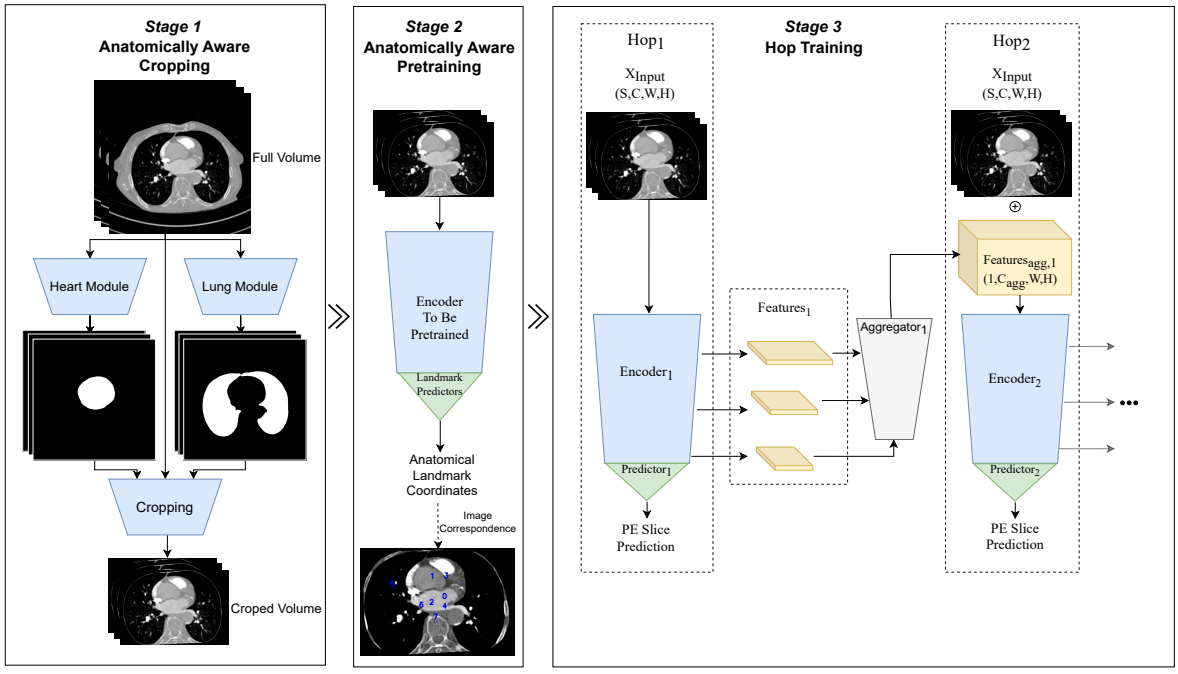

Figure 2.1: Proposed workflow: each stage represents one of the contributions. **Left**: Anatomically Aware Segmentation and Cropping, through which data is specialized for PE detection. **Middle**: Anatomically Aware Pretraining on the related task of Anatomical Landmark Detection, through which the model is primed for our task of Pulmonary Embolism detection. **Right** Hopped training, through which model predictions are refined, over two hops of neural processing.

Table 2.1: Comparison of results with state of the art solutions [1, 3, 40]. Our model, compared to state of the art [1, 3, 40], obtains better results in terms of F1 score. Our model performs slightly worse when compared to a human radiologist.

Model	**F1(↑)**	AUC PR(↑)
Weiker [40] AI Model	86.0 %	-
Cheikh [3] AI Model	86.1 %	-
Cheikh [3] Radiologist	*92.4 %*	-
Buls [1] AI model	73.0 %	-
Our baseline model	88.4 %	63.87 %
+ Cropping	89.0 %	64.29 %
+ Pretraining	89.4 %	65.95 %
+ Hop (Our final model)	**91.0 %**	**66.21 %**

level of 91.0%, outperforming other strong methods such as Weikert el al's [39] 86.0% F1 score. Full results are displayed in Table 2.1. The significant performance improvements over strong baselines and state-of-the-art methods demonstrate the effectiveness of our approach, even in the big data regime. Our model's performance, approaching that of expert radiologists, suggests its potential for clinical application.

Moreover, the three-stage framework introduced here represents a general proof of concept that could be adapted to other medical image analysis tasks. By breaking down complex diagnostic processes into intuitive, sequential steps, we've created a model that not only performs well but also aligns with human diagnostic reasoning.

Chapter 3

DEEP LEARNING FROM THE CLINICAL PERSPECTIVE

This study evaluated our deep neural network (DNN)-based algorithm for automated pulmonary embolism (PE) detection on CT pulmonary angiography (CTPA) scans. PE represents a serious condition ranking as the third most common cardiovascular syndrome worldwide, with prompt diagnosis being essential to prevent life-threatening complications. The algorithm was tested on 903 consecutive CTPA scans from a single university hospital, with a PE prevalence of 12.2% (n=110).

From a clinical perspective, PE diagnosis presents significant challenges. Radiologists face increasing workloads with CTPA scans growing 3-4% annually due to fear of missing PE, increased CT accessibility, and financial considerations [37]. In emergency settings, multiple life-threatening conditions must be ruled out alongside PE, including acute coronary syndrome, aortic dissection, and pneumonia. Oncologic patients face increased PE incidence due to malignancy-associated thrombosis, making early detection crucial for optimal management.

The DNN-based algorithm demonstrated strong performance with a sensitivity of 84.6%, specificity of 95.1% , and an accuracy of 93.8% for PE detection . This performance aligns with radiologists' typical sensitivity (67%-87%) and specificity (89%-99%) when consensus reading is considered the reference standard. Detailed results are presented in Table 3.1

This section is based on our work in: Tilman Emrich, Florin Condrea, et al "Accuracy of a Deep Neural Network for Automated Pulmonary Embolism Detection on Dedicated CT Pulmonary Angiograms"

Table 3.1: Overall Performance Analysis of the AI Model. 95% confidence intervals are in the squared brackets. All cases: n = 903.

Metric	Value	95% CI
True Positive	93	-
False Negative	17	-
False Positive	39	-
True Negative	754	-
Sensitivity	84.6%	[76.4%, 90.7%]
Specificity	95.1%	[93.3%, 96.5%]
Positive Predictive Value	70.5%	[63.5%, 76.6%]
Negative Predictive Value	97.8%	[96.6%, 98.6%]
Accuracy	93.8%	[92.0%, 95.3%]

Location-specific analysis revealed excellent sensitivity of 100% for central and of 96.7% for lobar PE, while peripheral PE detection was lower at 72.9% . This pattern mirrors clinical experience, where radiologists also demonstrate higher accuracy for central than peripheral PE (74%-86%) [19]. False negatives (n=17) primarily occurred with small chronic PEs in subsegmental arteries and acute PEs in segmental or subsegmental branches. False positives (n=39) were mainly caused by turbulent flow, artifacts, consolidation, and pulmonary masses. Detailed results are presented in Table 3.2.

Clinically, the algorithm performed similarly in obese and non-obese patients, demonstrating its robustness across different patient populations despite the challenges of reduced image quality in higher BMI patients due to increased noise, reduced spatial resolution, and beam hardening artifacts. The study population appropriately reflected real-world clinical scenarios, with PE-positive patients showing higher proportions of known risk factors including limited mobility, active malignancy, recent surgery, and clotting disorders.

These results suggest the algorithm's potential utility in supporting radiologists through exam prioritization and as a second reader, potentially improving workflow efficiency and diagnostic accuracy in clinical practice. Such AI integration could address the growing demand for CTPA interpretation while maintaining high diagnostic standards.

Table 3.2: Location-based Performance Analysis. 95% confidence intervals are in the squared brackets.

Location	True Positive	False Negative	Sensitivity
Overall	93	17	84.6% [76.4%, 90.7%]
Central	21	0	100.0% [83.9%, 100.0%]
Lobar	29	1	96.7% [82.8%, 99.9%]
Peripheral	43	16	72.9% [59.7%, 83.6%]

Chapter 4

WEAKLY SUPERVISED LEARNING VIA LABEL GRANULARITY ENCHANTMENT

In this chapter we present a weakly supervised learning pipeline called "Label Up" that upgrades annotation granularity from image-level to pixel-level for pulmonary embolism (PE) detection. The three-stage pipeline, displayed in Figure 4.1 first trains a slice-level classifier, then uses an explainability module based on integrated gradients [31] to generate pseudo-labels through iterative refinement, and finally trains a segmentation model on these pseudo-labels.

Our slice classifier, based on our previous work [7], achieves state-of-the-art performance with 97.1% AUCROC and 75.1% F1 score on slice level performance, providing a strong foundation for the explainability stage. The second stage employs an innovative iterative refinement process where predicted PE regions are progressively masked out, forcing the model to identify additional emboli. This approach significantly improves sensitivity by 10.1% while maintaining reasonable precision, with performance peaking after approximately six refinement iterations. The process automatically terminates when no further PEs are detected, optimizing computational efficiency. Post-processing filters based on size and anatomical location further enhance performance by eliminating false positives.

This section is based on our work in: Condrea, Florin, Saikiran Rapaka, and Marius Leordeanu. "Label up: Learning Pulmonary Embolism Segmentation from Image Level Annotation through Model Explainability." arXiv preprint arXiv:2412.07384 (2024).

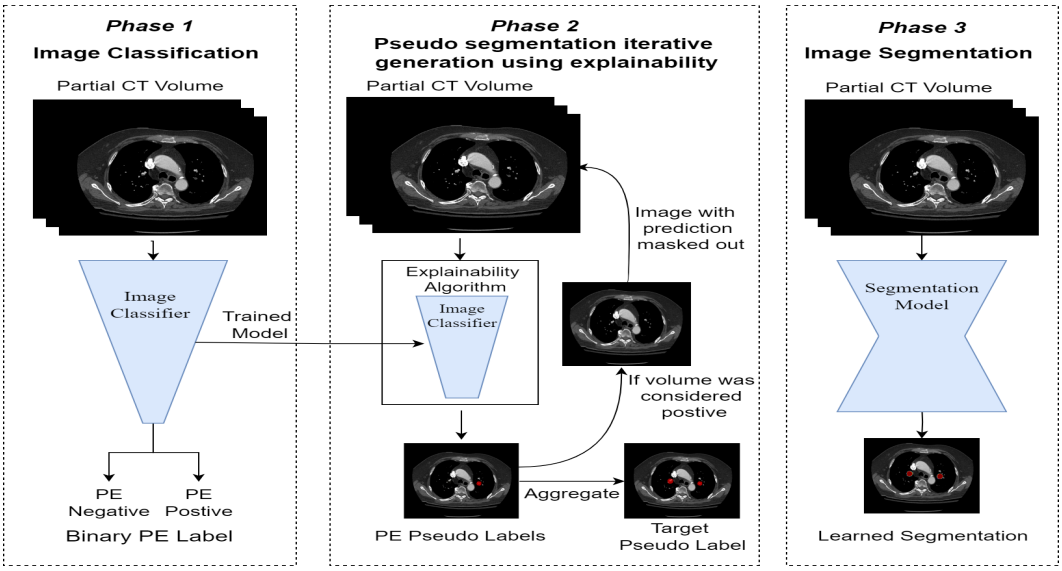

Figure 4.1: Our proposed pseudo-label multistage pipeline. In Stage 1, a slice level classifier is trained to predict if a stack of slices is positive. In Stage 2, the trained classifier, together with integrated gradients algorithm, predict the location of PEs. After each iteration, the predicted zones are masked out, and volume fed back into the network. Process repeats until PEs are no longer detected and all predictions are aggregated into pseudo-labels. In Stage 3, a deep learning segmentation network is trained on the pseudo-labels, obtaining performance superior to that of the explainability based teacher model.

When compared with other weakly supervised methods on PE localization, trained on a on a dataset of 6,824 studies [6] our approach achieves 69.4% F1 score on PE localization on a public dataset of 442 studies[2], slightly outperforming Pu et al.'s 69.1% F1 score which is trained on similar amounts of data[27] .By adding additional private data to our training dataset, obtaining a total of 13,329 studies, our performance improves to 71.6% F1 score, matching strongly supervised methods like Zhu et al. [42] despite using only image-level annotations. Full results are presented in Table 4.1.

Furthermore, finetuning our weakly supervised model on just 111 human-annotated studies yields 75.5% F1 score on PE localization, outperforming both random initialization (71.0%) and Zhu et al.'s method (71.6%) [42] by significant margins. This demonstrates that our weakly supervised pipeline not only generates high-quality pseudo-labels but also serves as an effective pretraining strategy when limited human annotations become available. Full results are presented in Table 4.2.

Our approach addresses a critical challenge in medical AI: the scarcity of fine-grained annotations. By leveraging readily available image-level labels to generate detailed segmentation masks, we enable the development of more precise diagnostic tools without the prohibitive cost of manual pixel-level annotation. The final segmentation model is also sub-

Table 4.1: Deep learning model performance. We report our performance on two datasets of 6,824 and 13,329 CTPA studies. Comparing with other pseudo-label methods [27] trained on a similar dataset, we obtain slightly better performance. By adding more private data to the training set, performance similar to strongly supervised methods [42] evaluated using the same detection matching criteria is obtained.

Model	**Data**	**Positives**	**Recall**	**PPV**	**F1**
Weakly Supervised Methods					
Pu et al [27]	6415	1990	61.8	78.2	69.1
Our Model	6824	1766	61.5	79.6	69.4
Our Model	13329	3389	66.9	77.0	71.6
Strongly Supervised Methods					
Ozkan et al [25]	142	142	95.1	52.6	67.7
Tajbakhsh et al [32]	121	121	83.4	47.2	60.3
Tajbakhsh et al [33]	121	121	32.9	98.6	49.4
Xu et al [41]	113	113	93.2	51.2	66.1
Zhu et al [42]	142	142	86.0	61.3	71.6
Weikert et al [40]	30000	15858	82.2	86.8	85.8

Table 4.2: Strongly supervised training results. We train our model on human annotation, starting from both random initialization and weakly supervised pretrained weights. Finetunning the model trained on pseudo-labels obtains a performance boost of 3.9 % F1 score, improving both sensitivity and PPV. The finetunned model also outperforms random initilization by 4.5 % F1 and another strongly supervised method trained on similar amount of data [42]. Our WSL model in italics is trained on pseudo-labels, and added for easier visualization of the improvement obtained through finetunning.

Model	**Data**	**Positives**	**Recall**	**PPV**	**F1**
Our WSL model Baseline	*13,329*	*3,389*	*66.9*	*77.0*	*71.6*
Zhu et al [42]	142	142	**86.0**	61.3	71.6
Our Random Init	111	111	66.2	76.6	71.0
Our WSL model Finetunned	111	111	73.9	**77.5**	**75.5**

stantially faster (50x) than the explainability-based teacher model while achieving better performance, making it suitable for clinical deployment. This methodology has broad potential applications beyond PE detection, offering a generalizable framework for upgrading annotation granularity across various medical imaging tasks.

Chapter 5

LEARNING TEMPORAL TASKS WITHOUT HUMAN ANNOTATIONS

In this chapter we explore an innovative approach to vital sign detection in thermal videos by leveraging synthetic data constructed from basic computer vision primitives. The research demonstrates how neural networks can recognize specific temporal signals despite significant differences between training and target domains. Our method, VSignNet, learns to predict both vital sign intensity and corresponding regions of interest in thermal videos without human supervision. The method achieves state-of-the-art results on the recent LCAS dataset [9].

Our approach consists of four stages: determining the frequency range of interest, generating synthetic training data, training the model, and performing inference with post-processing to detect signal peaks. The synthetic data generation methodology relies entirely on heuristic principles, creating training samples that share only essential target temporal signal characteristics. Our deep learning architecture captures temporal dimensions on two levels - local and global - enabling effective vital sign detection. As overview of out pipeline illustrated in Figure 5.1,

The VSignNet architecture processes input frames through a temporal convolutional encoder, followed by bidirectional LSTMs that aggregate global temporal information. Fully connected modules transform the embeddings into numerical values representing vital sign

This chapter is based on our work in: Condrea, Florin, Victor-Andrei Ivan, and Marius Leordeanu. "In search of life: Learning from synthetic data to detect vital signs in videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020.

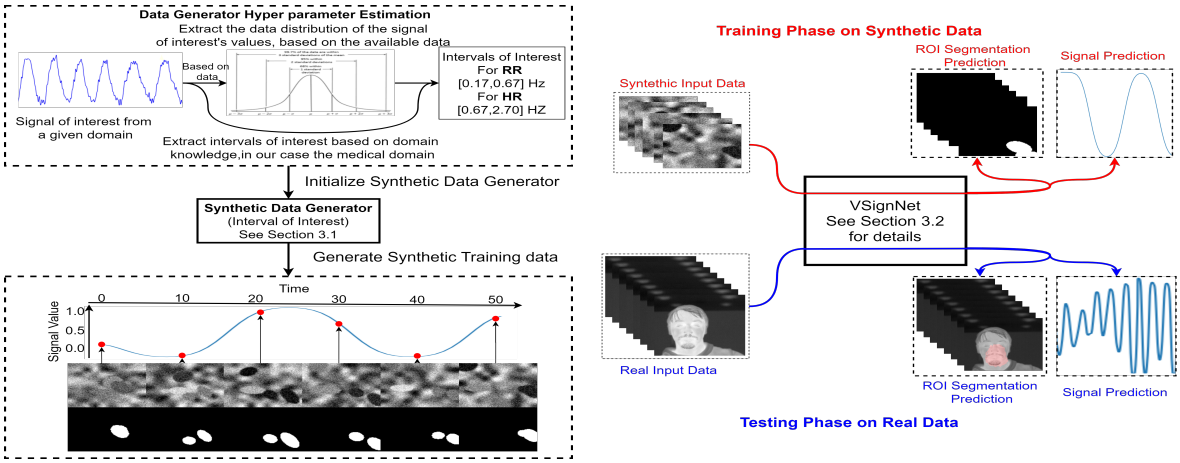

Figure 5.1: The four stages of our approach: the first stage consists of determining the frequency range of interest, either by using dataset statistics or domain knowledge. The second stage is the generation of synthetic training data samples. The third stage is the training of the model. At the final, inference stage, an additional post-processing step detects the peaks of the signal, on which we base our output average signal rate.

magnitude, while temporal convolutional decoders generate heatmaps encoding signal source locations. This dual-pathway approach allows for both signal intensity prediction and region-of-interest localization.

Heart Rate (BPM)				
Experiment	Still		Moving	
	MAE	STD	MAE	STD
LCAS [9]	29.68	±15.76	18.96	±22.51
VSignNet	**15.51**	± **9.93**	**14.91**	± **7.99**
Respiration Rate (BPM)				
Experiment	Still		Moving	
	MAE	STD	MAE	STD
LCAS [9]	3.72	±0.78	5.87	±2.18
M ROI	1.87	±2.05	4.41	±4.41
RF ROI	1.90	±1.72	14.77	±7.32
VSignNet	**1.12**	± **1.34**	**2.62**	± **2.07**

Table 5.1: Performance comparison between our results (VSignNet) and the results reported by Cosar et al (LCAS [9]). We also report the results of the two baselines, using either manually annotated ROI (M ROI) or ROI detected with RetinaFace (RF ROI).

VSignNet outperforms existing methods [5, 9, 23], achieving a mean absolute error of 1.12 (±1.34) for respiration rate per minute during still poses and 2.62 (±2.07) during movement, significantly better than the LCAS baseline (3.72±0.78 and 5.87±2.18 respectively). For heart rate detection, our method achieves a mean absolute error 15.51 (±9.93) of heartbeats per minute during still poses and 14.91 (±7.99) during movement, compared to LCAS's

29.68 (±15.76) and 18.96 (±22.51). The model also demonstrates robustness across different breathing patterns and effectively localizes regions of interest corresponding to the signal source. Full results are displayed in Table 5.1,

The remarkable efficacy of our relatively straightforward synthetic data generation algorithm in the complex domain of medical imaging merits further investigation. The successful application of such a methodology raises significant questions regarding the potential of synthetic training paradigms to facilitate unsupervised learning across diverse and complex spatiotemporal vision tasks [13, 22, 24]. This finding suggests promising avenues for addressing data scarcity challenges in specialized domains where annotated datasets are particularly difficult to obtain.

Chapter 6

Conclusions

This thesis has explored innovative approaches to address key challenges in medical AI, focusing on enhancing model performance under data constraints while maintaining clinical relevance. Our work has contributed to medical image analysis and vital sign monitoring through anatomically-informed design, synthetic data generation, and architectures leveraging both global and local features. We demonstrated the effectiveness of anatomically driven design in our HopNet framework, which utilized anatomical priors to guide preprocessing and model pretraining, resulting in more efficient feature learning and improved generalization.

To address data scarcity, we proposed two complementary approaches. Our Label-Up methodology increased annotation granularity without additional manual labeling by leveraging weak supervision and iterative refinement. In our vital sign monitoring work, we developed a method for generating synthetic training data based on target temporal dynamics, creating simplified geometric primitives that captured fundamental patterns for accurate respiratory and heart rate detection. Both approaches significantly reduced the annotation burden while maintaining high diagnostic accuracy.

Beyond technical innovations, we emphasized clinical validation and interpretability. Our collaboration with medical professionals provided insights into model strengths and limitations in realistic scenarios, particularly regarding challenging abnormalities. This clinically-informed evaluation methodology represents an important step toward developing AI systems that can meaningfully support medical decision-making in practice.

Bibliography

[1] Buls, N., Watté, N., Nieboer, K., Ilsen, B. and de Mey, J. [2021], 'Performance of an artificial intelligence tool with real-time clinical workflow integration–detection of intracranial hemorrhage and pulmonary embolism', *Physica Medica: European Journal of Medical Physics* **83**, 154–160.

[2] Callejas, M. F., Lin, H. M., Howard, T., Aitken, M., Napoleone, M., Jimenez-Juan, L., Moreland, R., Mathur, S., Deva, D. P. and Colak, E. [2023], 'Augmentation of the rsna pulmonary embolism ct dataset with bounding box annotations and anatomic localization of pulmonary emboli', *Radiology: Artificial Intelligence* **5**(3), e230001.

[3] Cheikh, A. B., Gorincour, G., Nivet, H., May, J., Seux, M., Calame, P., Thomson, V., Delabrousse, E. and Crombé, A. [2022], 'How artificial intelligence improves radiological interpretation in suspected pulmonary embolism', *European Radiology* pp. 1–12.

[4] Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K. and Ghassemi, M. [2021], 'Ethical machine learning in healthcare', *Annual Review of Biomedical Data Science* **4**, 123–144.

[5] Cho, Y., Julier, S. J., Marquardt, N. and Bianchi-Berhouze, N. [2017], 'Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging', *Biomedical Optics Express* **8**(10), 4480–4503.

[6] Colak, E., Kitamura, F. C., Hobbs, S. B., Wu, C. C., Lungren, M. P., Prevedello, L. M., Kalpathy-Cramer, J., Ball, R. L., Shih, G., Stein, A., Halabi, S. S., Altinmakas, E., Law, M., Kumar, P., Manzalawi, K. A., Nelson Rubio, D. C., Sechrist, J. W., Germaine, P., Lopez, E. C., Amerio, T., Gupta, P., Jain, M., Kay, F. U., Lin, C. T., Sen, S., Revels, J. W., Brussaard, C. C., Mongan, J., , Abdala, N., Bearce, B., Carrete, H., Dogan, H., Huang, S.-C., Crivellaro, P., Dincler, S., Kavnoudias, H., Lee, R., Lin, H.-M., Sale-

hinejad, H., Samorodova, O., Rodrigues dos Santos, E., Seah, J., Zia, A., Arteaga, V. A., Batra, K., Castelli von Atzingen, A., Chacko, A., DiDomenico, P. B., Gill, R. R., Hafez, M. A., John, S., Karl, R. L., Kanne, J. P., Mathilakath Nair, R. V., McDermott, S., Mittal, P. K., Mumbower, A., Lee, C., Orausclio, P. J., Palacio, D., Pozzessere, C., Rajiah, P., Ramos, O. A., Rodriguez, S., Shaaban, M. N., Shah, P. N., Son, H., Sonavane, S. K., Spieler, B., Tsai, E., Vásquez, A., Vijayakumar, D., Wali, P. P., Wand, A. and Zamora Endara, G. E. [2021], 'The rsna pulmonary embolism ct dataset', *Radiology: Artificial Intelligence* **3**(2), e200254. PMID: 33937862.
URL: *https://doi.org/10.1148/ryai.2021200254*

[7] Condrea, F., Rapaka, S., Itu, L., Sharma, P., Sperl, J., Ali, A. M. and Leordeanu, M. [2023], 'Anatomically aware dual-hop learning for pulmonary embolism detection in ct pulmonary angiograms', *arXiv preprint arXiv:2303.17593* .

[8] Condrea, F., Rapaka, S. and Leordeanu, M. [2024], 'Label up: Learning pulmonary embolism segmentation from image level annotation through model explainability', *arXiv preprint arXiv:2412.07384* .

[9] Cosar, S., Yan, Z., Zhao, F., Lambrou, T., Yue, S. and Bellotto, N. [2018], Thermal camera based physiological monitoring with an assistive robot, *in* 'IEEE International Engineering in Medicine and Biology Conference'.

[10] Davenport, T. and Kalakota, R. [2019], 'The potential for artificial intelligence in healthcare', *Future Healthcare Journal* **6**(2), 94–98.

[11] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Corrado, G., Thrun, S. and Dean, J. [2019], 'A guide to deep learning in healthcare', *Nature Medicine* **25**(1), 24–29.

[12] Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.-C., Kuo, P.-C., Lungren, M., Palmer, L., Price, B. J., Purkayastha, S., Pyrros, A., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., Trivedi, H., Wang, R., Zaiman, Z. and Zhang, H. [2022], 'Ai recognition of patient race in medical imaging: a modelling study', *The Lancet Digital Health* **4**(6), e406–e414.

[13] Gupta, A., Vedaldi, A. and Zisserman, A. [2016], Synthetic data for text localisation in natural images, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)'.

[14] He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X. and Zhang, K. [2019], 'The practical implementation of artificial intelligence technologies in medicine', *Nature Medicine* **25**(1), 30–36.

[15] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and M"uller, H. [2019], 'Causability and explainability of artificial intelligence in medicine', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **9**(4), e1312.

[16] Johnson, A. E. W., Pollard, T. J. and Mark, R. G. [2018], 'Reproducibility in critical care: a mortality prediction case study', *Proceedings of Machine Learning Research* **85**, 361–376.

[17] Kaissis, G. A., Makowski, M. R., R"uckert, D. and Braren, R. F. [2020], 'Secure, privacy-preserving and federated machine learning in medical imaging', *Nature Machine Intelligence* **2**(6), 305–311.

[18] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. [2019], 'Key challenges for delivering clinical impact with artificial intelligence', *BMC Medicine* **17**(1), 195.

[19] Kligerman, S. J., Mitchell, J. W., Sechrist, J. W., Meeks, A. K., Galvin, J. R. and White, C. S. [2018], 'Radiologist performance in the detection of pulmonary embolism: features that favor correct interpretation and risk factors for errors', *Journal of thoracic imaging* **33**(6), 350–357.

[20] Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. and Ferrante, E. [2020], 'Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis', *Proceedings of the National Academy of Sciences* **117**(23), 12592–12594.

[21] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B. and Sánchez, C. I. [2017], 'A survey on deep learning in medical image analysis', *Medical Image Analysis* **42**, 60–88.

[22] Mayer, N., Ilg, E., Fischer, P. et al. [2018], 'What makes good synthetic training data for learning disparity and optical flow estimation?', *International Journal of Computer Vision* **126**, 942–960.

[23] Mutlu, K., Esquivelzeta Rabell, J., Martin del Olmo, P. and Haesler, S. [2018], 'Ir thermography-based monitoring of respiration phase without image segmentation', *Journal of Neuroscience Methods* **301**, 1–8.

[24] Oh, T., Jaroensri, R., Kim, C., Elgharib, M., Durand, F., Freeman, W. and Matusik, W. [2018], 'Learning-based video motion magnification', *arXiv preprint arXiv:1804.02684* .

[25] Özkan, H., Osman, O., Şahin, S. and Boz, A. F. [2014], 'A novel method for pulmonary embolism detection in cta images', *Computer methods and programs in biomedicine* **113**(3), 757–766.

[26] Price, W. N. and Cohen, I. G. [2019], 'Privacy in the age of medical big data', *Nature Medicine* **25**(1), 37–43.

[27] Pu, J., Gezer, N. S., Ren, S., Alpaydin, A. O., Avci, E. R., Risbano, M. G., Rivera-Lebron, B., Chan, S. Y.-W. and Leader, J. K. [2023], 'Automated detection and segmentation of pulmonary embolisms on computed tomography pulmonary angiography (ctpa) using deep learning but without manual outlining', *Medical Image Analysis* **89**, 102882.

[28] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E. J. [2022], 'Ai in health and medicine', *Nature Medicine* **28**(1), 31–38.

[29] Ridnik, T., Ben-Baruch, E., Noy, A. and Zelnik-Manor, L. [2021], 'Imagenet-21k pre-training for the masses', *arXiv preprint arXiv:2104.10972* .

[30] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M. and Cardoso, M. J. [2020], 'The future of digital health with federated learning', *npj Digital Medicine* **3**(1), 119.

[31] Sundararajan, M., Taly, A. and Yan, Q. [2017], Axiomatic attribution for deep networks, *in* 'International conference on machine learning', PMLR, pp. 3319–3328.

[32] Tajbakhsh, N., Gotway, M. B. and Liang, J. [2015], Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks, *in* 'Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part II 18', Springer, pp. 62–69.

[33] Tajbakhsh, N. et al. [2019], 'Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and discriminative image representation', *Medical Image Analysis* **58**, 101541.

[34] Tan, M. and Le, Q. [2021], Efficientnetv2: Smaller models and faster training, *in* 'International Conference on Machine Learning', PMLR, pp. 10096–10106.

[35] Tjoa, E. and Guan, C. [2020], 'A survey on explainable artificial intelligence (xai)', *IEEE Access* .

[36] Topol, E. J. [2019], 'High-performance medicine: the convergence of human and artificial intelligence', *Nature Medicine* **25**(1), 44–56.

[37] Wang, R. C., Miglioretti, D. L., Marlow, E. C., Kwan, M. L., Theis, M. K., Bowles, E. J., Greenlee, R. T., Rahm, A. K., Stout, N. K., Weinmann, S. et al. [2020], 'Trends in imaging for suspected pulmonary embolism across us health care systems, 2004 to 2016', *JAMA network open* **3**(11), e2026930–e2026930.

[38] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R. M. [2020], 'Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases', *IEEE Transactions on Medical Imaging* **38**(7), 1588–1601.

[39] Weikert, T., Winkel, D. J., Bremerich, J., Stieltjes, B., Parmar, V., Sauter, A. W. and Sommer, G. [2020*a*], 'Automated detection of pulmonary embolism in ct pulmonary angiograms using an ai-powered algorithm', *European radiology* **30**, 6545–6553.

[40] Weikert, T., Winkel, D. J., Bremerich, J., Stieltjes, B., Parmar, V., Sauter, A. W. and Sommer, G. [2020*b*], 'Automated detection of pulmonary embolism in ct pulmonary angiograms using an ai-powered algorithm', *European radiology* **30**, 6545–6553.
URL: *https://pubmed.ncbi.nlm.nih.gov/32621243/*

[41] Xu, H., Li, H., Xu, Q., Zhang, Z., Wang, P., Li, D. and Guo, L. [2023], 'Automatic detection of pulmonary embolism in computed tomography pulmonary angiography using scaled-yolov4', *Medical Physics* **50**(7), 4340–4350.

[42] Zhu, H., Tao, G., Jiang, Y., Sun, L., Chen, J., Guo, J., Wang, N., Wei, H., Liu, X., Chen, Y. et al. [2024], 'Automatic detection of pulmonary embolism on computed tomography pulmonary angiogram scan using a three-dimensional convolutional neural network', *European Journal of Radiology* p. 111586.