PET and S-PET: A Two-Stage Pipeline for Pulmonary Embolism Detection from CTPA Scans

Anisha Raghu
Quarry Lane School
San Ramon, USA
anisharaghu10@gmail.com

Mihai Boicu
Information Sciences and Technology Department
George Mason University
Fairfax, USA
mboicu@gmu.edu

***Abstract*—Pulmonary embolism (PE) is a life-threatening condition with a high mortality rate of 30%, which requires immediate treatment for a positive outcome. The diagnosis of PE involves taking multiple X-ray images (often thousands of slices) of the chest, generating computed tomography pulmonary angiography (CTPA) images, and then analyzing them by a radiologist. However, the process often takes multiple days due to the limited availability of radiologists, suggesting the need for automation in the diagnosis process. In this research, diagnosis models are proposed based on the RadFusion dataset using CTPA scans and electronic health records (EHR) respectively. The labeling method and pre-processing of EHR data was improved and correlation-based analysis was performed to select key EHR features. A novel two-stage pipeline was developed to analyzing CTPA images. First, using a standard DinoV2 and Transformer architecture, a PE-Transformer model (PET) analyzes a window of contiguous CTPA slices (a chunk), followed by a Sequential PE-Transformer (S-PET) that aggregates information across multiple chunks. The PET model improved baseline accuracy results by 1.8% and the S-PET model by an additional 1.3%.**

Distilling the metadata features to just the 16 most important features and using a random forest classifier improved the AUROC from 0.76 to 0.79. The analysis supports the results from larger datasets such as INSPECT that showed that CTPA data is more informative than EHR data and bridges the gap in previous work. We also demonstrate that modern self-supervised backbones trained on web-scale data offer superior performance, reducing the need for custom architectures. This research also shows that very few EHR features contribute to accuracy, reducing the need to collect large amounts of EHR data. The current approach is not end-to-end trainable and separates chunk-level and patient-level models; in the future, we aim to explore unified models and more complex backbones.

***Index Terms*—Pulmonary Embolism Diagnosis, Computed Tomography Pulmonary Angiography Image Processing, Feature Selection, Electronic Health Records, Transformer Models, Random Forest Classifier**

I. Introduction

Pulmonary embolism (PE) is a life-threatening disease that is caused by a blockage in one of the pulmonary arteries, often caused by a blood clot traveling from another part of the body. This restricts blood flow to the lungs, causing serious complications in the ability of the heart to function properly [1]. PE contributes to more than 180,000 deaths annually in the United States [2], having a 30% mortality rate in undiagnosed and untreated cases. However, this can be reduced to eight percent with prompt diagnosis and treatment [3]. What exacerbates this condition is the difficulty in diagnosing this disease; the most common approach is using *(*computed tomography pulmonary angiography) (CTPA) images but image noise and poor opacification make this a difficult approach.

Deep learning applied to medical imaging, specifically CT scans, has shown significant promise in reducing the time it takes to diagnose critical diseases and accurately identify anomalies [4]–[6], with multiple studies trying to reduce the burden on radiologists in the diagnosis of PE [7]–[12]. CTPA scans, commonly used for PE diagnosis, are significantly larger than most medical imaging exams. The signs of pulmonary embolism show up only in a fraction of slices, making detection difficult [12]. Before the use of machine learning techniques, computer-aided detection (CAD) was used for the detection of PE in smaller, more curated datasets [13].

II. Related Research

A. Machine Learning Models

Several studies have explored the use of deep learning for pulmonary embolism (PE) detection and segmentation in CT imaging, as summarized in [14]. Ajmera et al. [15] employed a 2D U-Net architecture for clot segmentation, trained on a relatively small dataset of 251 patients. Their model achieved a sensitivity of 80%, specificity of 74%, and overall accuracy of 76%. Similarly, Liu et al. [16] trained a deep learning model on CT scans from 590 patients for clot localization. While the model achieved a high AUC of 0.92 and sensitivity of 94%, its specificity was limited to 76.5%, and the dataset size remained relatively small. A more large-scale approach was proposed by Weikert et al. [17], who trained a ResNet-based architecture on 28,000 CTA scans collected from multiple institutions. This model achieved strong performance, with a sensitivity

of 92.7% and specificity of 95.5%. However, a key limitation was that the test set was drawn from only a single institution, potentially limiting the model's generalizability [14].

Huang et al. introduced a custom model architecture (PENet) to automate the diagnosis of pulmonary embolism using CTPA scans using 3D convolutions to analyze a group of slices (chunk) to predict the probability of embolism for that chunk [12]. The probabilities are aggregated across chunks, and the maximum probability across chunks is used to predict PE for the patient. The results of this model are used as a baseline in this research. However, this work has a few limitations. Recent advances in self-supervised learning have led to the development of powerful general-purpose backbones reducing the need for custom model architectures [18], [19], [20]. In addition, machine learning techniques were not used to combine information across chunks.

The method outlined in [7] uses an ElasticNet model to predict PE from EHR data, reporting an AUROC of 93.2%.

B. Datasets

The most prominent datasets providing CTPA data are RSNA [21], RADFUSION [7] and INSPECT [11]. In addition, RadFusion and INSPECT provide EHR data. The RadFusion dataset was extensively studied in [12] and [7]. It consists of 1837 CT imaging studies, (comprising over 600000 2D slices), containing extensive EHR data (2853 metadata features) [7]. PE cases are categorized into three types: central, segmental, and subsegmental. The subsegmental cases are considered to have limited clinical value and are excluded from the dataset [12].

III. Experimental Design

A. Research Goals

Our research goals are as follows.

1) Investigate if general purpose pre-trained backbones leveraging self-supervised learning with a transformer encoder to aggregate information across slices outperform custom architectures.
2) Study whether better sampling of chunks and using a transformer encoder improve aggregating information across chunks.
3) Understand the relative importance of EHR versus CTPA images for detection models.

B. Data Labeling

In the RadFusion dataset, three labels were made for each patient by board certified radiologists: The presence or absence of PE, the type of PE and the slices were PE was present for cases with positive PE [7] [22]. In [7], [12] a patient was considered to have PE during training, only when slice level annotations were present. However, patient level labels were used for testing.

We propose a more consistent labeling approach considering a case positive only if both the patient level label and the slice level annotations are present for both training and testing. This corrects potential mislabeling at test, when samples with no slice level annotations are sometimes marked as positive. Table I displays the number of positive and negative cases after this correction, with the numbers reported in [7] in parentheses. Note that we use the same train/val/test split as defined in [7].

The change kept the training set balanced, but changed the validation and test sets to be unbalanced. We plan to address this in future research.

Category	Subset of Dataset			
	All	Train	Validation	Test
Negative Cases	1216 (1111)	946 (946)	136 (85)	134 (80)
Positive Cases	539 (726)	482 (508)	28 (108)	28 (110)

TABLE I
MODIFIED RADFUSION TABLE WITH SUBSEGMENTAL CASES REMOVED AND LABELS MODIFIED TO ALIGN WITH RADIOLOGISTS' OBSERVATIONS

C. Architecture

The architecture is based on a DinoV2 self-supervised Vision Transformer-small (ViT-S) and a transformer encoder to combine information between slices [18]. DinoV2 is a self-supervised vision model trained on more than 1 billion images to learn rich visual representations without manual labels, which is suitable for many downstream tasks. The ViT-Small architecture divides each image into patches and processes them through self-attention layers [23]. This backbone is well suited for medical imaging tasks because of its robustness and strong generalization. To capture contextual dependencies between multiple CT slices we used transformer encoder layers consisting of multi-head self-attention and feedforward layers [24]. To learn information across all slices into a single representation for classification we include a learnable class embedding at the beginning of the input sequence. Figure 1 describes the PE-Transformer (PET), which processes W=24 slices at a time and produces embeddings. The resulting embeddings are then concatenated and fed into a series of transformer encoder layers, followed by a linear layer, to produce the final embedding for each chunk. During training, the model learns to classify each chunk as positive or negative, similar to [12].

We use the chunk sampling strategy detailed in Figure 2, inspired by the observation that during inference, the prediction accuracy of the model depends on the absolute location of the slices where PE is visible. If the slices containing PE are spread across two chunks, the model would have lower confidence in detecting it. To remedy this, for a patient with W slices per N chunks, embeddings are generated for each chunk with an offset of 0, using the PET model, producing N embeddings. Then, we slide a window of size W with an offset of W/2, producing a second set of around N embeddings. These two sets - totaling around 2N embeddings - are concatenated and passed through a transformer encoder followed by a linear layer (Sequential PET model, Figure 3) to predict the probability that the patient has the disease. In contrast, [12] does not integrate information across chunks with a transformer; instead, it simply selects the maximum

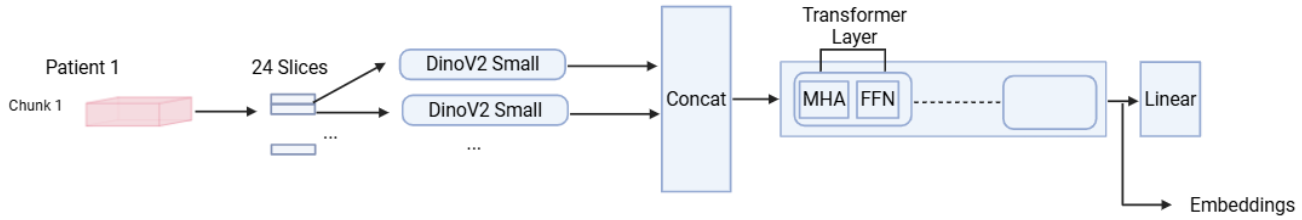

Fig. 1. Overview of PET Model consisting of DINOv2 + Transformer. The model processes one slice at a time for the W=24 slices in a chunk. The embedding produced by this model is processed by a transformer encoder and a linear layer for classification.

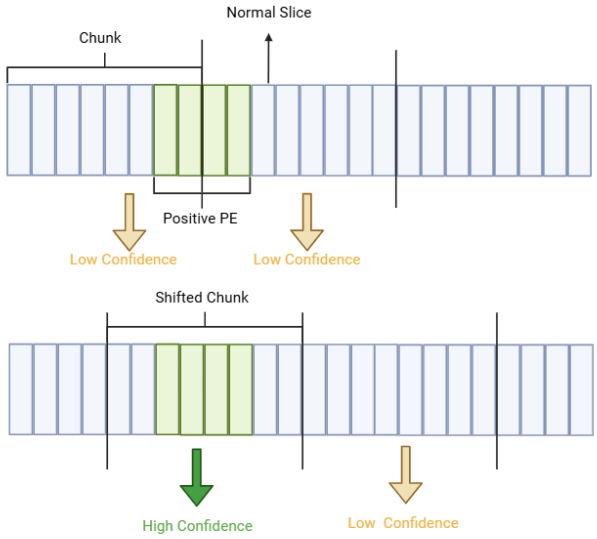

Fig. 2. Illustration of how shifting the starting offset of chunks can allow for a more accurate prediction.

probability from the N individual chunks. Our method not only generates twice as many embeddings but also captures inter-chunk relationships more effectively.

In addition to the data augmentation used in [12] we resized the images to a higher resolution - 256x256 compared to 196x196.

For the EHR data, we first discard all EHR features that do not show any variation in the training dataset, reducing about 50% of the features. We further rank the features by absolute value of the Pearson correlation coefficient with the disease labels and performed several experiments with different number of features. Our code and models are available at https://github.com/Anisha234/PET_SPET_Research.

IV. Results

A. EHR Only Model Results

We evaluated the performance of three classifiers—Decision Tree, Logistic Regression, and Random Forest—on the metadata features, both with and without feature selection based on correlation analysis. Table II summarizes the results across these two approaches on the best performing model, Random Forest. We note that the performance with EHR is poor, with the best approach having an AUC ROC of only 79.8% compared to using all the features. We note that using as few as 16 features shows improvements in all metrics except the accuracy and specificity.

B. Image Only Model Results

Compared to the Penet baseline with the modified labels, PET gives an improvement of 1.8% in AUROC and S-PET gives an additional improvement of 1.3% as shown in Table III.

To understand which components of PET and S-PET contribute most to performance, we ablate on image resolution, chunk sampling method and chunk fusion method Table IV . We see a gain of 0.7% by sampling chunks with an offset of half the chunk size. Using a transformer encoder for fusion further improves accuracy by 0.7%. Increasing the image resolution provides negligible improvement.

V. Conclusions

This work presents several important findings. First, label quality is a critical factor influencing model performance, and even small inconsistencies can lead to a significant changes in results. The improved performance of PET and S-PET architectures demonstrate that standard, high-capacity backbones can outperform custom models like Penet, even when the latter are specifically designed for the task of PE detection in CTPA scans. In addition, our results show that CTPA scans are more effective for PE diagnosis compared to EHR data, which aligns with [11]. We see multiple areas for further improvement: A limitation of our approach is that it is not end to end trainable. We train a model to predict probabilities/embeddings for a chunk (PET) and another model to predict the outcome for each patient (S-PET). A single stage approach that processes all the images would provide superior performance, at the cost of much higher complexity as we expect attention to dominate for long sequence lengths. In addition, the dataset used in this research was only obtained from one institution reducing generalizability and further evaluation of this model on datasets from different institutions would be desirable. In the future, we hope to experiment with more complex backbones and higher resolutions as we were limited by computational constraints.

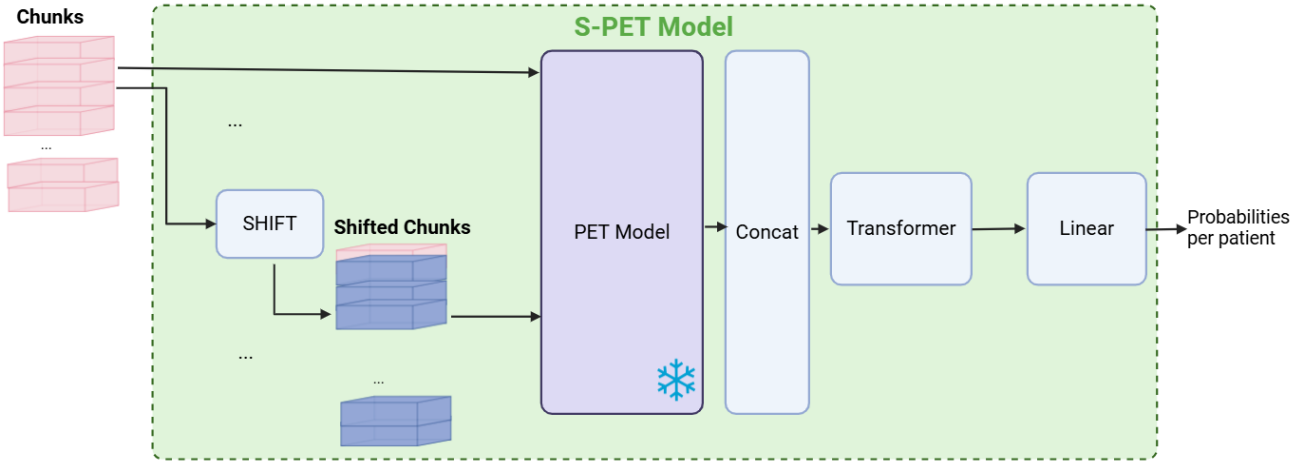

Fig. 3. Overview of S-PET Model: We first process all the chunks for a patient to produce embeddings for each chunk. This is also repeated with an offset of W/2. All the embeddings produced are concatenated and passed to a transformer encoder and linear layer to predict the probability of PE for a patient.

Classifier	Features	Sorted	Evaluation Metrics						
			Accuracy	Sensitivity	Specificity	Precision	F1	AUROC	Balanced Accuracy
Random Forest	2856	False	0.660	0.643	0.664	0.285	0.396	0.767	0.653
Random Forest	16	True	0.648	0.857	0.604	0.311	0.457	0.798	0.731

TABLE II

AN IMPROVEMENT OF ∼3.1% IN AUROC IS OBSERVED WHEN THE MOST RELEVANT METADATA FEATURES ARE USED INSTEAD OF ALL THE METADATA FEATURES. NOTE THAT EVEN WITH THIS IMPROVEMENT THE AUROC FOR THE EHR ONLY MODEL IS 79.8%.

Model	Evaluation Metrics						
	Accuracy	Sensitivity	Specificity	Precision	AUCROC	Balanced Accuracy	F1 Score
Penet Baseline	0.481	1.000	0.373	0.250	0.920	0.686	0.400
PET	**0.570**	**1.000**	**0.480**	**0.280**	**0.938**	**0.740**	**0.450**
S-PET	**0.796**	**0.892**	**0.776**	**0.454**	**0.951**	**0.834**	**0.733**

TABLE III

COMPARING THE PERFORMANCE OF PET AND S-PET WITH THE PENET BASELINE, THERE IS OVERALL AN IMPROVEMENT OF 3.1% PERCENT.

Image Size	Pooling Method	Include Shifted Chunks	Evaluation Metrics						
			Acc	Sensitivity	Specificity	Precision	AUROC	BA	F1
196	Max prob	False	0.48	1.00	0.37	0.35	0.937	0.68	0.40
256	Max prob	False	0.57	1.00	0.48	0.28	0.938	0.74	0.45
256	Max prob	True	0.55	1.00	0.45	0.28	0.944	0.73	0.43
256	**Transformer**	**True**	**0.79**	**0.89**	**0.78**	**0.45**	**0.951**	**0.83**	**0.73**

TABLE IV

A 1.7% IMPROVEMENT IN AUROC IS DUE TO AN IMPROVEMENT IN MODEL ARCHITECTURE—USING A STANDARD DINOV2SMALL + TRANSFORMER (PET) RATHER THAN THE PENET ARCHITECTURE (AUROC OF 92%). INCREASING THE RESOLUTION AND INTRODUCING CHUNK SAMPLING RESULTS IN A 0.7% INCREASE AND THE USE OF A TRANSFORMER FOR POOLING (S-PET) RATHER THAN CALCULATING THE MAX RESULTS IN A FURTHER 0.7% IMPROVEMENT.

VI. ACKNOWLEDGEMENTS

This research was made possible through the support of George Mason University's College of Science, which supports the ASSIP Program.

REFERENCES

[1] Cleveland Clinic, "Pulmonary embolism," https://my.clevelandclinic.org/health/diseases/17400-pulmonary-embolism, 2024, accessed: 2025-07-29.

[2] J. Soye, "Computed tomography pulmonary angiography: a sample of experience at a district general hospital," *The Ulster medical journal*, vol. 77, pp. 175–180, 2008.

[3] J. Bělohlávek, "Pulmonary embolism, part i: Epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism," *Experimental and clinical cardiology*, vol. 18, pp. 129–138, 2013.

[4] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. Patel, K. W. Yeom, K. Shpanskaya, S. Halabi, E. Zucker, G. Fanton, M. P. Lungren, and A. Y. Ng, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet," *PLoS Medicine*, vol. 15, no. 11, p. e1002699, 2018.

[5] J. J. Titano, M. Badgeley, J. Schefflein, M. Pain, A. Su, M. Cai, N. Swinburne, J. Zech, J. Kim, J. Bederson, J. Mocco, and E. K. Oermann, "Automated deep-neural-network surveillance of cranial images for acute neurologic events," *Nature Medicine*, vol. 24, pp. 1337–1341, 2018.

[6] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study," *The Lancet*, vol. 392, no. 10162, pp. 2388–2396, 2018.

[7] Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, and M. P. Lungren, "RadFusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR," Nov. 2021.

[8] H. Huhtanen, M. Nyman, T. Mohsen, A. Virkki, A. Karlsson, and J. Hirvonen, "Automated detection of pulmonary embolism from CT-angiograms using deep learning," *BMC Med. Imaging*, vol. 22, no. 1, p. 43, Mar. 2022.

[9] A. R. Hunsaker, "Deep learning and risk assessment in acute pulmonary embolism," *Radiology*, vol. 302, no. 1, pp. 185–186, Jan. 2022.

[10] A. M. Weng, A. E. Samir, A. Sharma, S. Lewis, V. Orhurhu, L. M. Prevedello, and M. K. Kalra, "Artificial intelligence for radiologic review of chest radiographs: Algorithm development and validation," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e210068, 2021.

[11] S.-C. Huang, Z. Huo, E. Steinberg, C.-C. Chiang, M. P. Lungren, C. P. Langlotz, S. Yeung, N. H. Shah, and J. A. Fries, "Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[12] S.-C. Huang, T. Kothari, I. Banerjee, C. Chute, R. L. Ball, N. Borus, A. Huang, B. N. Patel, P. Rajpurkar, J. Irvin, J. Dunnmon, J. Bledsoe, K. Shpanskaya, A. Dhaliwal, R. Zamanian, A. Y. Ng, and M. P. Lungren, "PENet-a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging," *NPJ Digit. Med.*, vol. 3, no. 1, p. 61, Apr. 2020.

[13] A. R. M. Al-Hinnawi, "Computer-aided detection, pulmonary embolism, computerized tomography pulmonary angiography: current status," in *Angiography*, B. Pamukçu, Ed. IntechOpen, 2018.

[14] P. A. Grenier, A. Ayobi, S. Quenet, M. Tassy, M. Marx, D. S. Chow, B. D. Weinberg, P. D. Chang, and Y. Chaibi, "Deep learning-based algorithm for automatic detection of pulmonary embolism in chest CT angiograms," *Diagnostics (Basel)*, vol. 13, no. 7, Apr. 2023.

[15] P. Ajmera, A. Kharat, J. Seth, S. Rathi, R. Pant, M. Gawali, V. Kulkarni, R. Maramraju, I. Kedia, R. Botchu, and S. Khaladkar, "A deep learning approach for automated diagnosis of pulmonary embolism on computed tomographic pulmonary angiography," *BMC Med. Imaging*, vol. 22, no. 1, p. 195, Nov. 2022.

[16] W. Liu, M. Liu, X. Guo, P. Zhang, L. Zhang, R. Zhang, H. Kang, Z. Zhai, X. Tao, J. Wan, and S. Xie, "Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning," *Eur. Radiol.*, vol. 30, no. 6, pp. 3567–3575, Jun. 2020.

[17] T. Weikert, D. J. Winkel, J. Bremerich, B. Stieltjes, V. Parmar, A. W. Sauter, and G. Sommer, "Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm," *Eur. Radiol.*, vol. 30, no. 12, pp. 6545–6553, Dec. 2020.

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," Apr. 2023.

[19] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," Jan. 2023.

[20] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," Nov. 2021.

[21] RSNA Pulmonary Embolism Detection Challenge, "Rsna str pulmonary embolism detection challenge dataset," https://www.kaggle.com/competitions/rsna-str-pulmonary-embolism-detection, 2020, accessed: 2025-07-29.

[22] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection," *Sci. Rep.*, vol. 10, no. 1, p. 22147, Dec. 2020.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," Oct. 2020.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Jun. 2017.