BILINGUAL PUBLISHING GROUP

Journal of Electronic & Information Systems

https://journals.bilpubgroup.com/index.php/jeis

ARTICLE

DPHT-Net: A Novel Self-Supervised Hybrid CNN-Transformer Approach for Automated Pulmonary Embolism Classification in CT Pulmonary Angiogram Scans

***Abeer Abdelhamid* [1,2*] , *Hossam El-Din Moustafa* [1,3], *Hala B. Nafea* [1], *Ehab H. Abdelhay* [1,4], *Mohammed M. Abo-Zahhad* [5], *Amir El-Ghamry* [6]**

[1] *Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt*

[2] *Higher Technological Institute of Applied Health Sciences, Mansoura 35511, Egypt*

[3] *Faculty of Artificial Intelligence and Information, Horus University, New Damietta 34517, Egypt*

[4] *Faculty of Engineering, Mansoura National University, Gamasa 35712, Egypt*

[5] *Electrical Engineering Department, Faculty of Engineering, Sohag University, Sohag 82524, Egypt*

[6] *Computer Science Department, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt*

ABSTRACT

Pulmonary embolism (PE) is a life-threatening condition that requires timely and accurate diagnosis to enable appropriate treatment and reduce mortality rates. Despite significant advances in medical imaging, reliable detection of PE in computed tomography pulmonary angiography (CTPA) remains challenging due to the subtle appearance of emboli, variations in contrast, and the high-dimensional nature of volumetric data. These challenges necessitate robust and efficient automated methods to support clinical decision-making. In this study, we propose a diffusion-pretrained hybrid Convolutional Neural Network (CNN)-Transformer network (DPHT-Net), a lightweight architecture designed to effectively capture both local and global embolic patterns in CTPA scans. The proposed framework integrates self-supervised diffusion-

*CORRESPONDING AUTHOR:

Abeer Abdelhamid, Electronics and Communications Engineering Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt; Higher Technological Institute of Applied Health Sciences, Mansoura 35511, Egypt; Email: beroabdo345@gmail.com

ARTICLE INFO

Received: 28 February 2026 | Revised: 13 April 2026 | Accepted: 20 April 2026 | Published Online: 25 April 2026
DOI: https://doi.org/10.30564/jeis.v8i1.13224

CITATION

Abdelhamid, A., Moustafa, H.E.-D., Nafea, H.B., et al., 2026. DPHT-Net: A Novel Self-Supervised Hybrid CNN-Transformer Approach for Automated Pulmonary Embolism Classification in CT Pulmonary Angiogram Scans. Journal of Electronic & Information Systems. 8(1): 97–108. DOI: https://doi.org/10.30564/jeis.v8i1.13224

based pretraining, a CNN-based module for local feature extraction and refinement, and a Transformer-based component for modeling long-range inter-slice dependencies. In addition, a sinh–cosh-based preprocessing step is introduced to enhance image contrast and highlight subtle embolic regions. The proposed model was evaluated on the RSNA-STR PE dataset. Experimental results demonstrate that DPHT-Net achieves an accuracy of 96.4% and an F1-score of 93.8%, outperforming conventional CNN-based methods by 11.9% in accuracy and 12.7% in F1-score, and surpassing Transformer-based approaches by 5.4% and 4.2%, respectively. These results indicate that DPHT-Net provides a robust, computationally efficient, and clinically applicable solution for automated PE detection, offering a promising direction for volumetric medical image analysis and computer-aided diagnosis systems.

Keywords: Artificial Intelligence (AI); Computed Tomography Pulmonary Angiograph (CTPA); Diffusion Pretrained Hybrid CNN-Transformer (DPHT); Pulmonary Embolism (PE)

1. Introduction

Pulmonary embolism (PE) is a fatal cardiovascular illness that accounts for the third greatest cause of cardiovascular death worldwide[1]. In clinical terms, PE is the third most prevalent cardiovascular emergency, behind heart attacks and strokes[2]. This pathology originates when blood clots clog the pulmonary arteries (PA), preventing blood flow and oxygen exchange in the lungs. A clot can form in the center of the main pulmonary artery or in its sub-branches, causing acute or chronic obstruction. This barrier reduces blood flow and damages the tissues[3]. The most common primary cause of PE is deep vein thrombosis (DVT), which usually begins as a thrombus in the deep veins of the lower limbs[4]. According to reports, death can reach 35% if therapy is not received; however, with prompt and adequate treatment, mortality can decrease to 2% to 15%[5]. Early diagnosis requires clinical examination, D-dimer tests, and radiological imaging[6]. Different imaging procedures can be used to diagnose PE, with computed tomography pulmonary angiography (CTPA) being the most common and highly regarded. Conversely, its use may result in undesirable consequences, particularly renal function impairment in people with a history of renal disease or allergies[7]. Furthermore, CTPA is time-consuming, costly, and unavailable in certain situations, particularly in grassroots healthcare facilities[8]. Current PE diagnostic techniques and tests have limitations, particularly in detecting segmental and subsegmental PEs, and are susceptible to inter-rater error[9]. As a result, developing an effective predictive model to identify high-risk PE patients prior to the commencement of clinical events is critical and significant.

Artificial intelligence (AI) has a significant impact on thromboembolic illnesses, particularly in terms of early prediction and diagnosis[10]. AI in healthcare will transform physician interpretation and diagnosis, ushering in a new era of medicine. Recent years have seen a significant surge in research investigations and scholarly papers evaluating the use of AI in medicine, including diagnostic radiology[11]. Deep learning (DL) methods, such as convolutional neural networks (CNNs) and recurrent neural networks, are particularly effective at processing medical imaging and time series data[12]. The use of multimodal DL models has the potential to reduce diagnostic mistake rates while also allowing for larger volumes in less time. Models using DL have shown enormous potential in the precise detection of PE due to their superior pattern recognition capabilities[13]. Natural language processing (NLP), a computerized method of parsing and extracting information from text, provides an automated, supplementary technique to phenotyping big datasets for quality review and clinical research[14].

Also, transformer language models are the most current version of NLP, and they are transforming the landscape in many industries, including healthcare[15,16]. They can analyze both the visible features and the location information of image patches, providing a more comprehensive view. They can use the self-attention process to consider the relationships between various PA image patches, ensuring the PA's continuity[17]. Human radiologists cannot review enormous amounts of imaging data as quickly as AI systems can. This enables faster diagnosis and treatment. Furthermore, AI may combine data from several sources, such as imaging studies, patient medical records, and laboratory testing, to provide a complete image of the patient's health status. To address

these issues, our research seeks to provide an AI-based platform specifically tailored for the accurate, efficient, and early diagnosis of PE using CTPA scans. The main contributions of this study are summarized as:

- Development of DPHT-Net, an efficient hybrid CNN–Transformer model for PE classification from volumetric CTPA scans.
- The proposed framework integrates self-supervised diffusion pretraining, CNN-based feature refinement, and Transformer-based inter-slice aggregation to effectively capture both local and global embolic representations.
- Experimental results on the RSNA-STR PE dataset demonstrate that the proposed model achieves 96.4% accuracy superior, outperforming state-of-the-art CNN and transformer-based methods.

2. Related Work

The use of AI approaches has been shown to improve the diagnostic accuracy of many imaging modalities. This includes detecting, identifying, and quantifying illness conditions. Condrea et al.[18] trained their model in segments, utilizing a dual-hop deep neural network as the framework. The model was trained on CTPA scans from 12,012 patients in the RSPECT dataset to identify anatomical features as regions of interest. The sensitivity of 92% and specificity of 96.1% show that PE detection is accurate. Adding a step of training to boost the model's anatomical awareness improved overall performance. Ayobi et al.[19] assessed CINA-PE, an FDA-approved AI application for detecting PE during CTPA. The suggested technique showed great sensitivity and specificity of 93.3% and 94.8%, respectively. Khan et al.[20] used DenseNet201 as a deep learning framework to diagnose PE using CAD. The model categorizes images into nine categories: negative PE, indeterminate PE, right-sided PE, left-sided PE, central PE, RV/LV ratio > 1, RV/LV ratio < 1, acute PE, and chronic PE. The results indicated an accuracy of 88% [0.86–0.90], sensitivity of 88% [0.86–0.90], specificity of 89% [0.87–0.91], and AUC of 0.90 [0.88–0.92].

Also, Suman et al.[21] presented a similar pipeline for the RESPECT dataset and evaluated their model on a curated external dataset, where the AUROC of positive studies reached 0.949. Yang et al.[22] used a two-stage CNN, including a candidate method and a false positive eradication subnet. It had a sensitivity of 75.4%. Ma et al.[23] suggested a two-phase multitask learning algorithm that can detect the presence of PE and its properties, such as the acute or chronic location and the related RV/LV ratio, lowering false-negative diagnoses. Their method scored an AUROC of 0.93 and a sensitivity of 86% on RSNA-STR dataset. Ting et al.[24] employed 3D CNN model to classify PE from CTPA images. Their model achieved an accuracy of 85% and AUROC of 0.84 on their dataset. Biret et al.[25] developed HybridNeXt architecture which combined traditional CNN models with the Swin transformer for PE identification from CTPA images. Their technique reported an accuracy of 90.14%.

Additionally, Singh et al.[26] proposed CNN-LSTM classifier model for PE classification task. Their model trained on RSNA-STR dataset and reported an AUC of 0.82. da Silva et al.[27] suggested a hybrid model combined with InceptionResNetV2 with RNN-LSTM models for PE classification from CTPA scans. Their model achieved an accuracy of 93% on RSNA-STR dataset. Huang et al.[28] presented a 77-layer 3D CNN model called PENet, to detect PE from CTPA images. Their technique achieved an AUROC of 0.84 on their dataset. Kiourt et al.[29] presented a DL-based method for localizing and categorizing PE from CTPA images. Their method is based on YOLO V4 model and achieved 91.6% accuracy. Amudha and Sunitha[30] explored a late-fusion strategy for multimodal PE classification. The late fusion strategy involved training distinct DL models independently on image and EMR data, then combining their outputs at the decision level to preserve modality-specific learning. Their technique was evaluated on the Stanford University Medical Center (SUMC) dataset and scored an accuracy of 90.8%. A model called EmbNet was presented by Zhu et al.[31] for automatic PE detection from CTPA scans. Their approach achieved sensitivity between (83.5–86%).

Despite the promising performance of recent DL models for PE classification, several important limitations remain. CNN-based architectures, such as ResNet and Xception, are effective in capturing local spatial features but fail to model long-range inter-slice dependencies in volumetric CTPA data. Transformer-based approaches improve global contextual modeling; however, they typically require large annotated datasets and involve high computational complexity, which limits their practical applicability. Furthermore, most existing methods rely heavily on fully supervised learning,

restricting their ability to generalize in scenarios with limited labeled medical data and subtle embolic patterns. These limitations highlight the need for more efficient and robust representation learning strategies.

To address these challenges, the proposed DPHT-Net integrates self-supervised diffusion pretraining with a hybrid CNN-Transformer architecture. This design enables effective local feature extraction, global context modeling, and improved data efficiency, thereby overcoming key limitations of recent state-of-the-art methods.

3. Materials and Methods

In 2020, the Radiological Society of North America (RSNA) and the Society of Thoracic Radiology (STR) held their fourth ML competition, the RSNA-STR PE detection challenge[32]. This dataset includes over 12,000 CTPA examinations from five international research facilities, with expert annotations contributed by over 80 thoracic radiologists. Images in a CT examination were sorted sequentially based on image location coordinates found in the DICOM header, which was critical for a competitive solution. We used the publicly available RSNA-STR dataset to develop the proposed model. In this work, the task is formulated as scan-level binary classification (PE vs. Normal), where each CTPA examination is assigned a single label. **Figure 1** represents the structure of the proposed DPHT-Net model. The model starts with the preprocessing of raw DICOM volumes to homogeneous inputs and improves embolic features. After turning each scan into a grayscale image, lung windowing is applied with a window level of 600 HU and a window width of 1,500 HU, improving pulmonary vasculature contrast. From the 3D volume, 64–80 axial slices on each side, the lungs are extracted. All the slices will be resized to 256 × 256 pixels and normalized to the range [0,1]. A sinh–cosh intensity transformation is applied as a standard preprocessing step to enhance image contrast according to the following equation:

$$I_{\text{processed}} = \sinh(I_{\text{normalized}}) + \cosh(I_{\text{normalized}}) \quad (1)$$

$I_{\text{normalized}}$ is followed by rescaling to maintain numerical stability. This transformation is adopted as a standard contrast enhancement technique and is not intended as a methodological contribution, but rather as a supporting preprocessing step. Massive data augmentation is done, such as random rotation ($\pm 20^\circ$), flipping, and adjustments of brightness/contrast ($\pm 20\%$). The preprocessed slices are then input to a self-supervised diffusion denoising autoencoder with diffusion-inspired training[33], where a CNN-based encoder maps each slice x_i into a latent vector $z_i \in \mathbb{R}^{128}$:

$$z_i = f_{\text{encoder}}(x_i), \quad (2)$$

and a UNet-style denoising diffusion decoder g_{decoder} is trained to reconstruct the input from a noise-perturbed latent representation $\tilde{z}_i$ using a score-matching loss $L_{\text{diffusion}}$[34]:

$$L_{\text{diffusion}} = \mathbb{E}_{x_i, \epsilon \sim \mathcal{N}(0,1)} \left[\| x_i - g_{\text{decoder}}(z_i + \epsilon) \|_2^2 \right] \quad (3)$$

Each CT slice was first processed through a lightweight CNN-based encoder within a self-supervised diffusion autoencoder, producing a 128-dimensional latent vector. This CNN efficiently captures slice-level spatial and textural features, including subtle embolic patterns, while maintaining computational efficiency. Following the diffusion pretraining, the latent vectors were refined via a lightweight CNN head employing depthwise separable convolutions, batch normalization, and ReLU activations, further enhancing the discriminative power of each slice representation. The sequence of refined slice vectors $\{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_N\}$ was then fed into an encoder-only Transformer with multi-head self-attention. Each slice vector acts as a token, allowing the Transformer to capture inter-slice dependencies across the entire scan. The output sequence was aggregated via attention pooling to form a single scan-level vector, which was subsequently passed through fully connected layers for binary classification of PE. This architecture leverages the CNNs for local slice-level feature extraction and the Transformer for global scan-level context modeling, enabling accurate and interpretable exam-level predictions. After pretraining, encoder weights are frozen, producing a sequence of latent vectors $[z_1, z_2, \ldots, z_N]$ for all slices, preserving spatial and contextual information. Each latent vector is refined by a lightweight CNN head:

$$\hat{z}_i = \text{ReLU}\big(\text{BN}(\text{SepConv}(z_i))\big), \quad (4)$$

where $\text{SepConv}(\cdot)$ represents depthwise separable convolution and $\text{BN}(\cdot)$ denotes batch normalization. Positional encoding $\text{PE}(i)$ is added to retain slice order:

$$\tilde{z}_i = \hat{z}_i + \text{PE}(i), \quad i = 1, \ldots, N. \quad (5)$$

The sequence $\{\tilde{z}_i\}_{i=1}^{N}$ is processed by a Transformer encoder with multi-head self-attention[35]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (6)$$

where $Q, K, V \in \mathbb{R}^{N\times d_k}$ are the query, key, and value matrices, and d_k is the attention dimension[36]. Attention pooling aggregates the sequence into a single scan-level vector z_{scan}:

$$z_{\text{scan}} = \sum_{i=1}^{N} \alpha_i \tilde{z}_i \quad (7)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{N}\exp(e_j)}, \quad e_i = \text{score}(\tilde{z}_i) \quad (8)$$

where α_i represents the attention weight of slice i. The scan-level vector is passed through fully connected layers with ReLU activations and dropout to produce the final probability of PE:

$$\hat{y} = \sigma\left(W_2\,\text{ReLU}(W_1 z_{\text{scan}} + b_1) + b_2\right) \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function, and W_1, W_2, b_1, b_2 are learnable parameters. The network is trained using binary cross-entropy loss[37]:

$$L_{\text{BCE}} = -\frac{1}{M}\sum_{i=1}^{M}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right] \quad (10)$$

using AdamW optimizer, 1×10^{-4} initial learning rate, cosine learning rate decay, and warm-up. The stepwise workflow of the proposed DPHT-Net model is illustrated in **Algorithm 1**. To conclude, DPHT-Net is an integration of self-supervised diffusion pretraining, CNN-based refinement of features, and Transformer-based aggregation. The sinh–cosh preprocessing enhances image contrast, while the diffusion-based encoder learns strong slice-level representations, and the Transformer exploits inter-slice dependencies, which enables accurate scan-level classification while preserving a degree of interpretability through attention-based slice weighting and visualization.

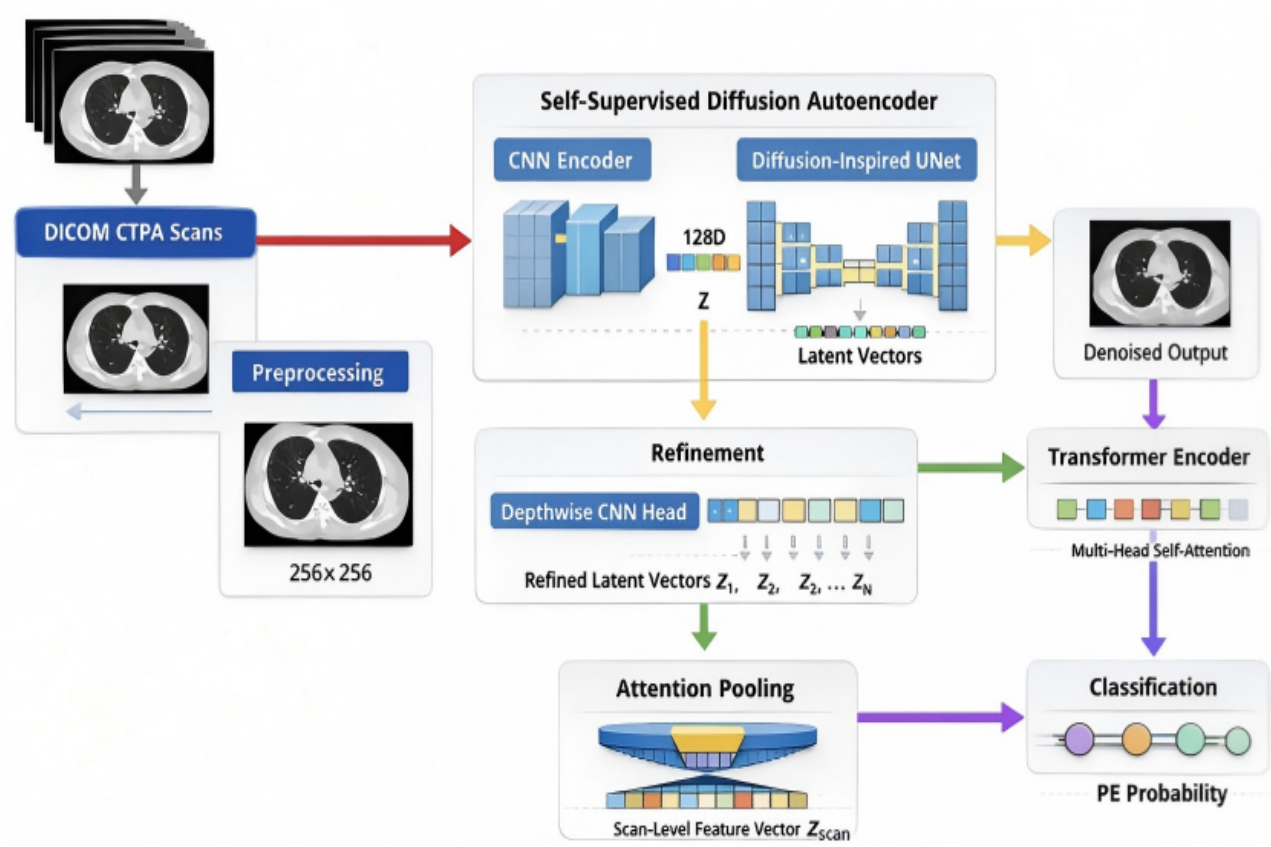


Figure 1. Overview of the proposed DPHT-Net model.

Algorithm 1 DPHT-Net Workflow for PE Classification from CTPA Scans

Require: Raw CTPA scans $\{Scan_i\}_{i=1}^{M}$
Ensure: PE probability prediction $\hat{y}$ for each scan
function DPHT-NET($Scan$)
  for all slice x_i in $Scan$ **do**
    $x_i \rightarrow$ ConvertToGrayscale(x_i)
    $x_i \rightarrow$ LungWindow($x_i, WL = -600, WW = 1,500$)
    $x_i \rightarrow$ Resize($x_i, 256, 256$)
    $x_i \rightarrow$ Normalize($x_i, 0, 1$)
    $x_i \rightarrow$ Sinh-CoshTransform(x_i)
    $x_i \rightarrow$ Data Augmentation(x_i)
  end for
  for all slice x_i in $Scan$ **do**
    $z_i \rightarrow$ CNNEncoder(x_i)
    $\hat{z}_i \rightarrow$ CNNHead(z_i)
    $\tilde{z}_i \rightarrow \hat{z}_i +$ Positional Encoding(i)
  end for
  $Z_{scan} \rightarrow$ Transformer Encoder($\{\tilde{z}_i\}_{i=1}^{N}$)
  $z_{scan} \rightarrow$ Attention Pooling(Z_{scan})
  $\hat{y} \rightarrow$ Sigmoid(FC Layers(z_{scan}))
  return $\hat{y}$
end function
Optimize BCE Loss with AdamW, cosine LR decay, warm-up

4. Results

Classification of PE remains challenging, as subtle embolic features are not always discernible on CTPA scans, but early diagnosis is important. Here, we have proposed DPHT-Net, a hybrid CNN-Transformer model, with diffusion-based pretraining, for automatic PE diagnosis. The performance of the model was evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC curve, providing a comprehensive assessment of its discriminative ability. A stratified 5-fold cross-validation scheme was used, and the final results are reported as the mean across all folds. We present a comprehensive analysis of its performance against traditional models and state-of-the-art models in **Table 1**, along with an ablation study demonstrating the effect of each component (**Table 2**). The performance of DPHT-Net was compared with traditional CNN-based architectures and modern Transformer-based methods. As shown from **Table 3**, Resnet50, a traditional CNN network reached an accuracy of 84.5%, with 80.2% precision, 82.0% recall and 81.1% F1-score. Depthwise separable convolutions based Xception reached a slightly higher accuracy of 88.0%. Further hierarchical Transformer-based Swin Transformer increased accuracy to 91.0%.

Table 1. Previous studies for PE classification using different AI/DL methods.

Study	Model	Dataset	Performance Metrics
Condrea et al. [18]	Dual-hop deep neural network	RESPECT (12,012 CTPA scans)	Sensitivity: 92%, Specificity: 96.1%
Ayobi et al. [19]	CINA-PE (FDA-approved AI)	Private dataset (CTPA scans)	Sensitivity: 93.3%, Specificity: 94.8%
Khan et al. [20]	DenseNet201	CAD images	Accuracy: 88%, Sensitivity: 88%, Specificity: 89%, AUC: 0.90
Suman et al. [21]	Pipeline similar to RESPECT	RESPECT + external dataset	AUROC: 0.94
Yang et al. [22]	Two-stage CNN	Private dataset (CTPA scans)	Sensitivity: 75.4%
Ma et al. [23]	Two-phase multitask learning	RSNA-STR	AUROC: 0.93, Sensitivity: 86%
Ting et al. [24]	3D CNN	Private dataset (CTPA images)	Accuracy: 85%, AUROC: 0.84

Table 1. *Cont.*

Study	Model	Dataset	Performance Metrics
Biret et al. [25]	HybridNeXt (CNN + Swin Transformer)	Private dataset (CTPA images)	Accuracy: 90.14%
Singh et al. [26]	CNN-LSTM	RSNA-STR	AUC: 0.82
da Silva et al. [27]	Hybrid Inception-ResNetV2 + RNN-LSTM	RSNA-STR	Accuracy: 93%
Huang et al. [28]	77-layer 3D CNN (PENet)	Private dataset (CTPA images)	AUROC: 0.84
Kiourt et al. [29]	YOLO V4	Private dataset (CTPA images)	Accuracy: 91.6%
Amudha and Sunitha [30]	Late fusion (image + EMR)	SUMC dataset	Accuracy: 90.8%
Zhu et al. [31]	EmbNet	Private dataset (CTPA scans)	Sensitivity: 83.5–86%
DPHT-Net (Proposed)	**DPHT-Net (Hybrid pretrained CNN-Transformer architecture)**	**RSNA-STR**	**Acc = 96.4%, F1-score = 93.8%**

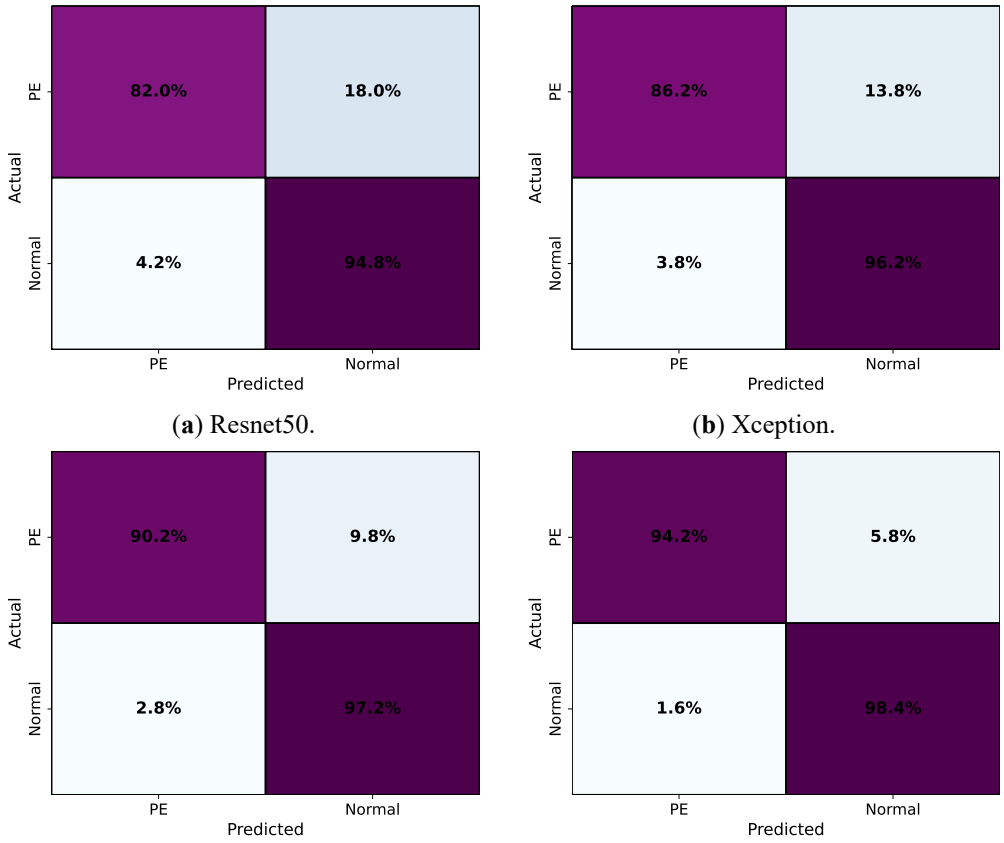
Table 2. Ablation study of the proposed DPHT-Net components on PE CTPA dataset.

Variant	Acc. (%)	Pre. (%)	Rec. (%)	F1-Score (%)
DPHT-Net without Transformer	91.0	88.5	89.2	88.8
DPHT-Net without Diffusion	92.5	89.0	90.0	89.5
DPHT-Net without CNN refinement	93.2	90.0	90.8	90.4
DPHT-Net (Proposed)	96.4	93.5	94.2	93.8

Table 3. Performance comparison of the proposed model with traditional models.

Model	Acc. (%)	Pre. (%)	Rec. (%)	F1-Score (%)
ResNet50	84.5	80.2	82.0	81.1
Xception	88.0	85.0	86.2	85.6
Swin Transformer	91.0	89.0	90.2	89.6
DPHT-Net (Proposed)	96.4	93.5	94.2	93.8

The proposed DPHT-Net significantly outperformed all baseline models, achieving 96.4% accuracy, 93.5% precision, 94.2% recall, and 93.8% F1-score. These results highlight the model's superior ability to detect subtle embolic features across volumetric CTPA scans, outperforming both traditional CNNs and recent Transformer-based approaches. To provide a clearer quantitative comparison, the relative performance gains of the proposed DPHT-Net over baseline models are explicitly analyzed. Specifically, DPHT-Net improves accuracy by 11.9% compared to ResNet50, 8.4% compared to Xception, and 5.4% compared to the Swin Transformer. Similarly, the F1-score shows notable improvements, confirming the robustness of the proposed model across different evaluation metrics. These results highlight the effectiveness of the hybrid architecture in capturing both local and global features in volumetric CTPA scans. To understand the contribution of each component in DPHT-Net, we performed an ablation study (**Table 2**). Removing the Transformer encoder caused the largest drop in performance, reducing accuracy to 91.0%. Excluding diffusion pretraining resulted in an accuracy of 92.5%, while omitting the CNN refinement led to 93.2% accuracy. The full DPHT-Net consistently achieved the best results (96.4% accuracy), confirming the contribution of each architectural component. Additionally, **Figures 2** and **3** display the confusion matrices and the combined ROC curve comparison of the proposed model against other traditional CNN and Transformer models. Also, to enhance model interpretability, attention maps were generated to visualize the regions contributing most to the model's predictions. As illustrated in **Figure 4**, the model exhibits non-uniform attention patterns, focusing on localized regions that are more informative for classification. This behavior suggests that the model prioritizes relevant image features rather than relying on global or spurious correlations, thereby improving transparency and supporting its clinical applicability.



(**a**) Resnet50. (**b**) Xception.

(**c**) Swin transformer. (**d**) Proposed (DPHT-Net).

Figure 2. Confusion matrices depicting the performance of the proposed model against other traditional models.

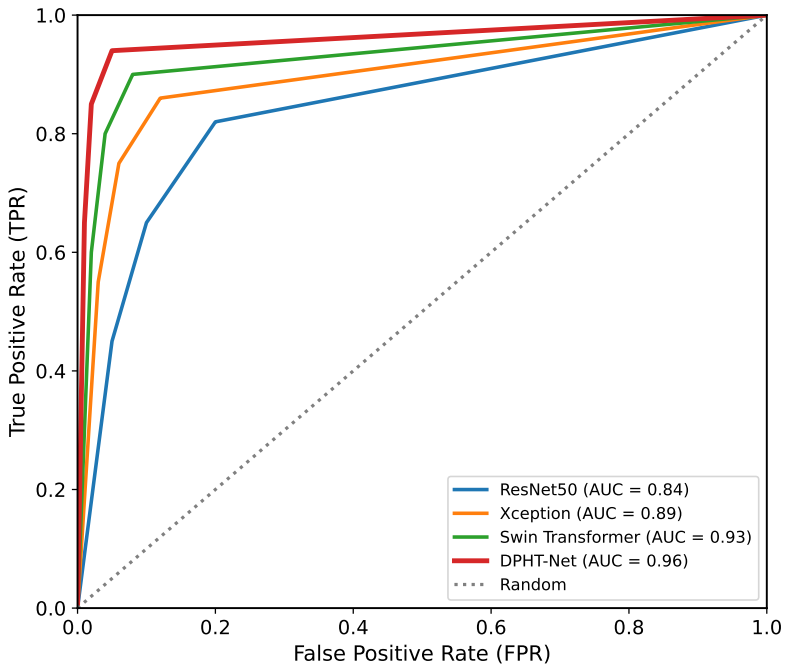


Figure 3. ROC curve comparison of ResNet50, Xception, Swin Transformer, and the proposed DPHT-Net on the RSNA dataset.
Note: The figure illustrates the superior discriminative performance of DPHT-Net, achieving the highest AUC compared to all baseline and transformer-based models, indicating improved classification capability for PE classification.

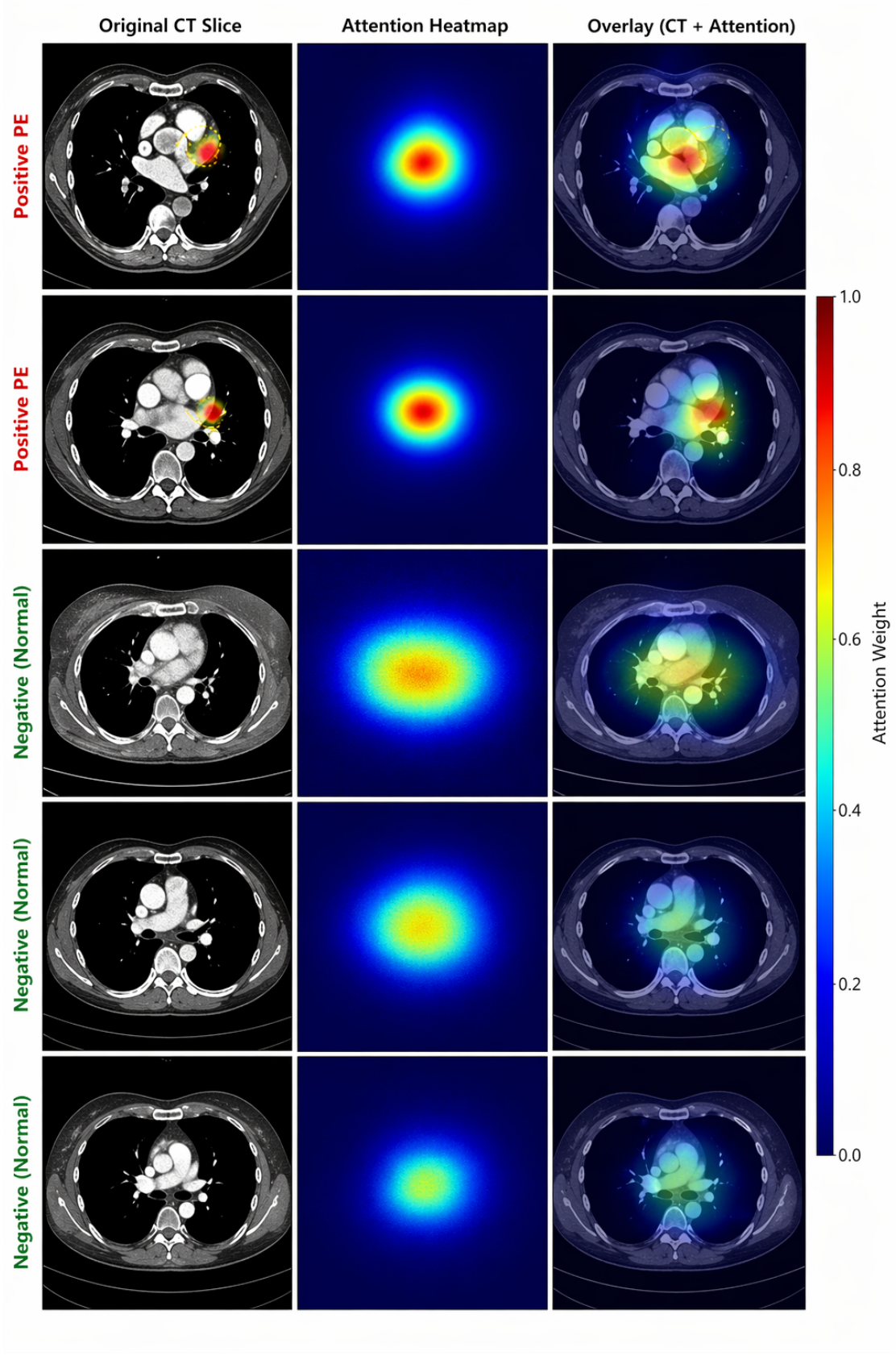


Figure 4. Attention maps illustrating model attention for positive and negative cases.

5. Discussion

The experimental results demonstrate that DPHT-Net provides a highly effective framework for PE classification in CTPA scans. By integrating diffusion pretraining, CNN-based feature refinement, and Transformer-based inter-slice aggregation, DPHT-Net consistently outperforms both traditional CNN architectures and recent Transformer-based methods. The achieved accuracy of 96.4% and F1-score of 93.8% underscore the model's ability to capture subtle embolic features that are challenging for simpler architectures.

Comparison with baseline models highlights several key insights. As shown from **Table 3** and **Figures 2** and **3**, a more detailed analysis reveals that the performance gains of DPHT-Net are not merely due to increased model complexity, but rather the complementary interaction between its architectural components. Specifically, the diffusion-based pretraining stage enables the model to learn noise-robust latent representations, which is particularly important in CTPA scans where embolic regions often exhibit low contrast and high variability. This leads to more stable and discriminative feature embeddings compared to conventional supervised training.

In addition, the CNN-based refinement module enhances local spatial feature discrimination by leveraging depthwise separable convolutions, allowing the model to better capture fine-grained embolic patterns within individual slices. This is especially beneficial for detecting small or peripheral emboli that may be overlooked by standard CNN architectures. Furthermore, the Transformer encoder plays a critical role in modeling inter-slice dependencies across the volumetric scan. Unlike traditional CNN-based approaches that process slices independently, the proposed method captures contextual relationships between consecutive slices, enabling more accurate identification of spatially distributed embolic structures.

These observations are further supported by the ablation study, where removing the Transformer component resulted in the most significant performance degradation, confirming its importance in capturing global contextual information. Similarly, excluding diffusion pretraining led to a noticeable drop in performance, highlighting its contribution to robust feature learning.

The ablation study clearly demonstrates the contribution of each component as obtained in **Table 2**. Removing the Transformer module caused the largest drop in accuracy to 91.0%, confirming its critical role in modeling inter-slice dependencies. Excluding diffusion pretraining, CNN refinement, or sinh–cosh preprocessing led to gradual performance declines, showing that each component complements the others to maximize effectiveness. This stepwise evaluation underscores that the hybrid design of DPHT-Net is essential for achieving state-of-the-art results in PE classification, as summarized in **Table 1**.

Compared with existing CNN-only and Transformer-only methods, DPHT-Net achieves a better balance between local feature sensitivity and global contextual modeling. This hybrid design, combined with self-supervised pretraining, improves generalization and enhances the detection of subtle embolic patterns. As illustrated in **Figure 4**, the attention maps provide insight into how the model aggregates information across slices, offering a degree of transparency in the decision-making process. By assigning higher weights to specific slices, the model emphasizes the most informative regions within the scan. Qualitative analysis of these attention patterns suggests that the model focuses on localized and clinically relevant structures rather than uniformly processing all slices or relying on background artifacts. This behavior supports the assumption that the model captures meaningful embolic-related features that contribute to the final prediction.

Despite strong quantitative results, several limitations remain. The dataset, while representative, may not capture the full variability of PE presentation across diverse populations and imaging protocols. Real-time clinical deployment may require further optimization, and external validation on larger, multi-center datasets would strengthen generalizability and robustness.

6. Conclusions

In this study, we introduced DPHT-Net, a hybrid CNN-Transformer model for automated PE classification in CTPA scans. By combining self-supervised diffusion pretraining, CNN-based feature refinement, and Transformer-based inter-slice aggregation, DPHT-Net effectively captures subtle embolic patterns that are often missed by conventional architectures. Evaluation results demonstrate that DPHT-Net outperforms traditional CNNs and recent Transformer-based

models, achieving an accuracy of 96.4%, and an F1-score of 93.8%. The ablation study confirms that each component contributes meaningfully to the overall performance, highlighting the importance of the hybrid design in achieving robust and accurate PE classification.

For future work, we plan to extend DPHT-Net to larger multi-center datasets to ensure generalizability across different populations and imaging protocols. Additionally, integrating patient clinical information alongside imaging features could further improve predictive performance and enhance its potential for real-world clinical deployment. Moreover, model optimization techniques such as lightweight compression and efficiency-aware design may be explored to facilitate real-time deployment in resource-constrained clinical settings.

Author Contributions

All authors contributed to the study's conception and experimental design. Material preparation and data collection: A.A., A.E.-G., E.H.A., and H.E.-D.M.; Data analysis, validation and visualization: A.A., A.E.-G., M.M.A.-Z., and H.B.N.; Supervision: H.E.-D.M., A.E.-G., E.H.A., H.B.N., and M.M.A.-Z.; Software: A.A., A.E.-G.; Writing—original draft preparation: A.A. and A.E.-G.; Writing—reviewing and editing: A.A., A.E.-G., H.E.-D.M., H.B.N., and M.M.A.-Z. All authors commented, edited, and approved the final manuscript before submission.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The dataset used is publicly available at https://www.kaggle.com/c/rsna-str-pulmonary-embolism-detection/data.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Uzelac, B., Stanković, S., 2026. Artificial Intelligence-Driven Integration of ECG and Molecular Biomarkers in Pulmonary Embolism. International Journal of Molecular Sciences. 27(2), 813.

[2] Rivas, L.F., 2023. Clinical characterization of patients with venous thromboembolic disease in 2 reference centers in El Salvador. Blood. 142(Supplement 1), 5555.

[3] Nitha, V.R., Vinod Chandra, S.S., Valsalan, P., et al., 2025. A deep learning framework for the segmentation and quantitative analysis of pulmonary embolism. Engineering Applications of Artificial Intelligence. 155, 110972.

[4] Abdelhamid, A., El-Ghamry, A., Abdelhay, E.H., et al., 2025. Improved pulmonary embolism detection in CT pulmonary angiogram scans with hybrid vision transformers and deep learning techniques. Scientific Reports. 15(1), 31443.

[5] Douillet, D., Roy, P.-M., Penaloza, A., 2021. Suspected acute pulmonary embolism: gestalt, scoring systems, and artificial intelligence. Seminars in Respiratory and Critical Care Medicine. 42(2), 176–182.

[6] Mohanarajan, M., Salunke, P.P., Arif, A., et al., 2025. Advancements in Machine Learning and Artificial Intelligence in the Radiological Detection of Pulmonary Embolism. Cureus. 17(1).

[7] Williams, L.-M.S., Walker, G.R., Loewenherz, J.W., et al., 2020. Association of contrast and acute kidney injury in the critically ill: A propensity-matched study. Chest. 157(4), 866–876.

[8] Zhou, Q., Huang, R., Xiong, X., et al., 2025. Prediction of pulmonary embolism by an explainable machine learning approach in the real world. Scientific Reports. 15(1), 835.

[9] Bass, A.R., Fields, K.G., Goto, R., et al., 2017. Clinical decision rules for pulmonary embolism in hospitalized patients: A systematic literature review and meta-analysis. Thrombosis and Haemostasis. 117(11), 2176–2185.

[10] Stamate, E., Piraianu, A.-I., Ciobotaru, O.R., et al., 2024. Revolutionizing cardiology through artificial intelligence—Big data from proactive prevention to precise diagnostics and cutting-edge treatment—A comprehensive review of the past 5 years. Diagnostics. 14(11), 1103.

[11] Beam, A.L., Drazen, J.M., Kohane, I.S., et al., 2023. Artificial intelligence in medicine. New England Journal of Medicine. 388(13), 1220–1221.

[12] Rajkomar, A., Dean, J., Kohane, I., 2019. Machine learning in medicine. New England Journal of Medicine. 380(14), 1347–1358.

[13] Huhtanen, H., Nyman, M., Mohsen, T., et al., 2022. Automated detection of pulmonary embolism from CT-angiograms using deep learning. BMC Medical Imaging. 22(1), 43.

[14] Lam, B.D., Ma, S., Kovalenko, I., et al., 2025. Using a transformer language model to curate a pulmonary embolism dataset from the Medical Information Mart for Intensive Care IV: MIMIC-IV-Ext-PE. Research and Practice in Thrombosis and Haemostasis. 9(4), 102896.

[15] Omiye, J.A., Gui, H., Rezaei, S.J., et al., 2024. Large language models in medicine: the potentials and pitfalls: a narrative review. Annals of Internal Medicine. 177(2), 210–220.

[16] Abdelhaliem, I., Dixon, J., Abdelhamid, A., et al., 2025. A Multimodal Adaptive Inter-Region Attention-Guided Network for Brain Tumor Classification. IEEE Access. 13, 187964–187975.

[17] Qiao, Y., Gao, Y., Chen, Y., et al., 2025. Quantitative assessment and risk stratification of random acute pulmonary embolism cases using a deep learning model based on computed tomography pulmonary angiography images. Quantitative Imaging in Medicine and Surgery. 15(3), 1950.

[18] Condrea, F., Rapaka, S., Itu, L., et al., 2024. Anatomically aware dual-hop learning for pulmonary embolism detection in CT pulmonary angiograms. Computers in Biology and Medicine. 174, 108464.

[19] Ayobi, A., Chang, P.D., Chow, D.S., et al., 2024. Performance and clinical utility of an artificial intelligence-enabled tool for pulmonary embolism detection. Clinical Imaging. 113, 110245.

[20] Khan, M., Shah, P.M., Khan, I.A., et al., 2023. IoMT-enabled computer-aided diagnosis of pulmonary embolism from computed tomography scans using deep learning. Sensors. 23(3), 1471.

[21] Suman, S., Singh, G., Sakla, N., et al., 2021. Attention based CNN-LSTM network for pulmonary embolism prediction on chest computed tomography pulmonary angiograms. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; pp. 356–366.

[22] Yang, X., Lin, Y., Su, J., et al., 2019. A two-stage convolutional neural network for pulmonary embolism detection from CTPA images. IEEE Access. 7, 84849–84857.

[23] Ma, X., Ferguson, E.C., Jiang, X., et al., 2022. A multitask deep learning approach for pulmonary embolism detection and identification. Scientific Reports. 12(1), 13087.

[24] Ting, I.-H., Tseng, Y.-J., Lin, Y.-S., 2026. Application of deep learning techniques in non-contrast computed tomography pulmonary angiogram for pulmonary embolism diagnosis. arXiv preprint. arXiv:2601.00925.

[25] Biret, C.B., Gurbuz, S., Akbal, E., et al., 2025. Advancing Pulmonary Embolism Detection with Integrated Deep Learning Architectures. Journal of Imaging Informatics in Medicine. 39, 186–201.

[26] Singh, G., Singh, A., Kainth, T., et al., 2025. Comparing efficiency of an attention-based deep learning network with contemporary radiological workflow for pulmonary embolism detection on CTPA: A retrospective study. European Journal of Radiology Open. 14, 100657.

[27] da Silva, L.O., da Silva, M.C.B., Ribeiro, G.A.S., et al., 2024. Artificial intelligence-based pulmonary embolism classification: Development and validation using real-world data. Plos One. 19(8), e0305839.

[28] Huang, S.-C., Kothari, T., Banerjee, I., et al., 2020. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. NPJ Digital Medicine. 3(1), 61.

[29] Kiourt, C., Feretzakis, G., Dalamarinis, K., et al., 2021. Pulmonary embolism identification in computerized tomography pulmonary angiography scans with deep learning technologies in COVID-19 patients. arXiv preprint. arXiv:2105.11187.

[30] Amudha, T.K., Sunitha, R., 2025. Multi-Modal Deep Learning for Pulmonary Embolism Classification: A Late Fusion Approach Combining CTPA Imaging and EMR Data. In proceedings of the 2025 International Conference on Emerging Technologies in Engineering Applications (ICETEA), Puducherry, India, 5–6 June 2025.

[31] Zhu, H., Tao, G., Jiang, 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. 177, 111586.

[32] Colak, E., Kitamura, F.C., Hobbs, S.B., et al., 2021. The RSNA pulmonary embolism CT dataset. Radiology: Artificial Intelligence. 3(2), e200254.

[33] Ho, J., Jain, A.N., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems. 33, 6840–6851.

[34] Song, Y., Sohl-Dickstein, J., Kingma, D.P., et al., 2020. Score-based generative modeling through stochastic differential equations. arXiv preprint. arXiv:2011.13456.

[35] Vaswani, A., Shazeer, N., Parmar, N., et al., 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.

[36] Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. Proceedings of the 35th International Conference on Machine Learning. 80, 2127–2136.

[37] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press: Cambridge, MA, USA.