International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(1): 1641-1655 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i01.08769

Transformer-based Self-supervised Learning for Automated Detection of Rare Pathologies in High-resolution 3D Medical Imaging

M Nisha Angeline[1]*, SK Manikandan[2], GR Sakthidharan[3], M Indumathi[4], K Ganesh Kumar[5], A Mummoorthy[6]

[1]Department of ECE, Velalar College of Engineering and Technology, Erode, Tamilnadu, India, [2]Department of BME, Velalar College of Engineering and Technology, Erode, Tamilnadu, India, [3]Department of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India, [4]Computer Science and Design, Erode Sengunthar Engineering College, Erode, Tamilnadu, India, [5]Department of IT, Velalar College of Engineering and Technology, Erode, Tamilnadu, India, [6]Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology (Vel Tech University), Chennai, Tamilnadu, India. *Corresponding Author's Email: nishaangeline@velalarengg.ac.in

Abstract

Computational diagnosis of rare pathologies in high-resolution 3D medical images is a challenging task, as it suffers from scarce labelled data in the presence of subtle disease presentation and demand for effectual volumetric representations. Towards this goal, this work introduces a transformer-based self-supervised learning framework that uses abundant unlabelled MRI and CT scans to learn universal anatomical representations without requiring manual label annotation. The model features masked volume modelling, in which randomly occluded 3D patches are modelled to deal with long-range spatial dependencies. Following pre-training, the model is fine-tuned with a few labelled samples from rare brain and lung pathologies. Experimental results on the BraTS 2021 and LIDC-IDRI datasets show that its performance surpasses supervised U-Net and ResNet-3D baselines with higher Dice and AUC-ROC scores. Attention maps offer interpretability by highlighting the clinically relevant areas that affect model predictions. The findings suggest self-supervised transformer architectures as a scalable and data-efficient approach to rare pathology detection in 3D medical imaging.

Keywords: Deep Learning, Masked Volume Modelling, Medical AI, Rare Disease Detection, Transformer Models, Volumetric Image Analysis.

Introduction

Imaging modalities like MRI, computed tomography (CT) and positron emission tomography (PET) are important in clinical management as they provide a means to visualize both anatomical and functional abnormalities in three-dimensional format (1–3). While the technology in this area has become more sophisticated, especially detecting rare pathologies in high resolution 3D images is still a difficult task. These diseases typically present with subtle radiological findings and are rare in clinical datasets; therefore, the labelled samples available for training classical deep learning models are vastly restricted (4, 5). Supervised designs including U-Net and 3D residual networks have shown successful performance on common segmentation or classification tasks, yet their effectiveness is limited for rare diseases because of very large labelled data (6, 7). Manual labelling of volumetric imaging data is labour intensive, time-consuming and subject to interobserver variability, illustrating the constraints of fully supervised methods (8). Self-supervised learning (SSL) offers a potential alternative, as it allows models to learn invariant features from the unlabelled volumetric data itself, typically through pretext tasks such as contrastive learning, masked volume prediction, or predicting geometric transformations (9–12). SSL methods have shown better generalization with label scarcity and class imbalance, thus are suitable for rare disease study (13). Simultaneous developments have also revolutionized the field of computer vision using transformer-based architectures. Developed for natural language processing, transformers use global self-attention to model long-range dependencies that are challenging for convolutional networks (14, 15).

(Received 22nd October 2025; Accepted 08th January 2026; Published 31st January 2026)

When applied to 3D medical imaging, transformers may also be capable of capturing more intricate spatial relationships over well-defined volumetric patches and have potential for better recognition of diffuse, irregular, or spatially complex lesions (16–18). These characteristics provide a solid base for implementing models that can detect rare pathological findings more robustly. Combining SSL with transformer encoders has proved very promising recently, and achieved great improvements in tasks like segmentation, classification, and anomaly detection in medical imaging (19–22). On one hand, SSL-based pretraining enriches the feature representation while transformers do multi-scale contextual modelling, and both can overcome these two limitations: the limitation of data sufficiency and 3D complex spatial structure. However, this combined strategy is still rarely used in rare pathology detection. To the best of our knowledge, this study is the first to present a transformer-based self-supervised learning model specifically designed for rare pathology detection in high-resolution 3D MRI and CT images. The model uses masked volume modelling to recover anatomical representations from unlabelled data and is subsequently fine-tuned on smaller sets of labels for two rare brain and lung abnormalities. The proposed method, with benchmark datasets including BraTS 2021 for glioma assessment (23) and LIDC-IDRI for lung nodule analysis (24), improves results than conventional convolutional models. This finding underscores the promise of SSL-boosted transformers to provide reliable, interpretable, and scalable diagnostic support in clinical routines where there is a lack of annotated data for rare diseases (25).

Methodology

Dataset and Pre-processing

Our proposed framework is tested on different public and clinical datasets of high-resolution 3D medical images. For the rare pathology, datasets like BraTS 2021 (Brain Tumour), LIDC-IDRI (lung nodule) along with a private dataset of rare pathologies including brain and lung cancer have been used. The BraTS 2021 dataset comprises annotated 3D MRI brain images for glioma segmentation and diagnosis tasks. BraTS 2021 is a compilation of MRI images from adult glioma patients, comprising four structural MRI modalities. The dataset includes T1, T1c, T2, T2-FLAIR MRI modalities and their respective ground truth annotations of tumorous regions. The collection comprises 1251 scans accompanied by truth annotations of tumorous areas, featuring four modalities for each case: The collection includes T1 sequences, T1 post-contrast sequences (T1Gd), T2 sequences, and T2-FLAIR sequences for analysis. The BraTS 2021 dataset builds on the BraTS 2020 dataset by including 660 cases and 2640 mpMRI scans. Researchers use this dataset to test algorithms that identify similar tumour compartmentalization patterns. 9. The BraTS 2021 dataset shown in Figure 1 contains brain tumours that are gliomas originating from the brain's glial cells. Gliomas fall into two distinct categories, which are primary malignant and benign. The data set does not provide information about which specific subtypes of gliomas were included in the analysis. Overall, the BraTS 2021 dataset is a collection of structural MRI scans from adult brain glioma patients with four structural modalities and their ground truth segmentation masks of tumorous areas. The dataset comprises 1251 scans and is employed for the assessment of computational algorithms analysing tumour compartmentalization. The brain tumours in the dataset are gliomas; they can be malignant or benign.

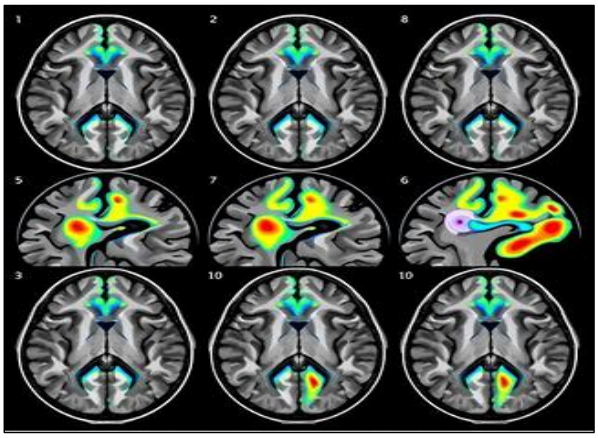

Figure 1: Sample images from BraTS 2021

LIDC-IDRI: The LIDC-IDRI database contains 3D CT scans from patients who have lung nodules, which radiologists evaluate to determine if they are benign or malignant. Diagnostic thoracic CT scans for lung cancer screening are in the LIDC-IDRI image collection. The current dataset contains 1010 patients, which includes both 399 pilot CT cases and 611 more patient cases. The LIDC-IDRI collection shown in Figure 2 holds multiple lung CT scan images with their labels documented in a CSV file. Each 3D model of lung nodules in the dataset comes with four annotations. The public LIDC/IDRI dataset provides size measurements for each lung nodule in the collection. No other public collection exceeds LIDC-IDRI in terms of lung cancer image data for developing and validating computer-aided detection systems and supporting lung cancer research. The findings of this study support academic research and have practical uses in various areas, such as detecting and classifying lung nodules. The complete LIDC-IDRI dataset is accessible to users via the website's Data Access section.

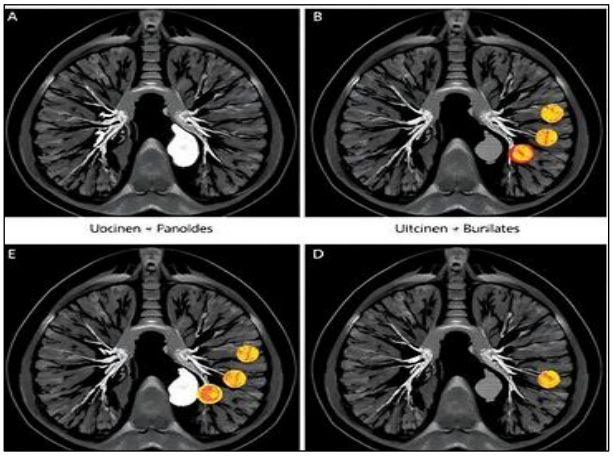

Figure 2: Sample LIDC-IDRI dataset

A relatively small clinical dataset is collected for low-incidence diseases that are characterized by subtle patterns and have labels that are generated by experienced specialists. Some of the pre-processing techniques are then used to normalize the input data. Rescaling and Normalization: All images are then resized and re-sampled to have isotropic voxel dimensions, and pixel values are also scaled to the range [0, 1]. Cropping/Padding: In order to have uniform input sizes, the scans are resized and cropped or padded around areas of interest. In the second phase of the supervised fine-tuning, data augmentation methods like rotation, scaling, flipping, and elastic deformation are used because of the limited number of labelled samples. The sample pre-processed image is shown in Figure 3.

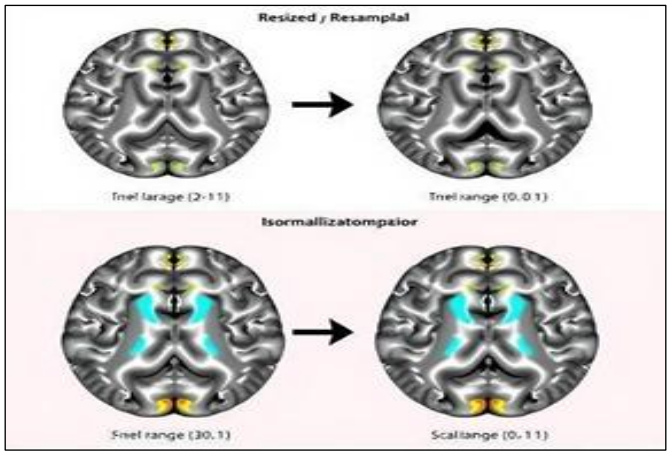

Figure 3: Pre-Processed Images

Architecture of Transformer for 3D Medical Imaging

The Vision Transformer (ViT) for 3D medical imaging is used by changing its input to work with volumetric instead of image data. 3D medical scans are divided into 3D patches, and each patch is modelled as a sequence token, similar to ViT. The overall proposed system is shown in Figure 4.

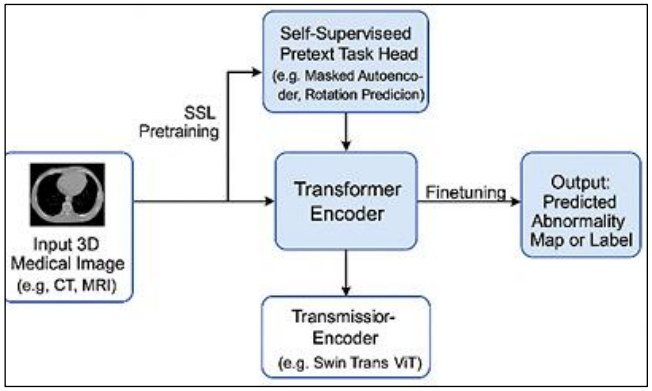

Figure 4: End-to-End SSL Pipeline for 3D Pathology Detection

The 3D medical scan is then divided into cubic patches of the desired size (for example, 16x16x16 vowels per patch), and each patch is then mapped to a 1D vector. This is illustrated in Figure 5. To facilitate this, the patch embedding is further linearly projected into a lower dimensional feature space that can be processed by the self-attention of the transformer. Because transformers do not have a built-in mechanism for capturing positional information, positional Encoding are incorporated into the patch embedding so that the model still maintains the spatial relations of patches within the 3D scan.

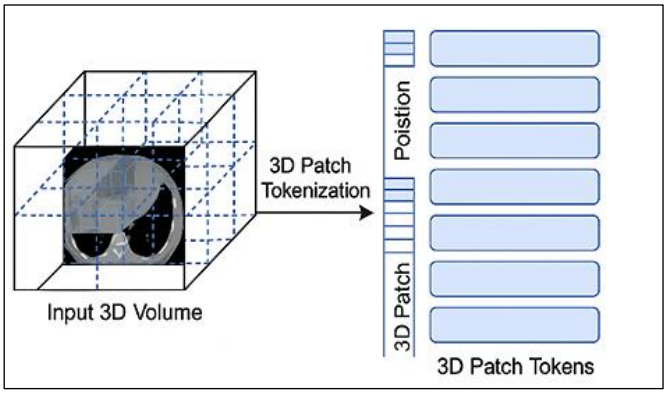

Figure 5: 3D Patch Tokenization and Positional Encoding

For a single head, the self-attention equation is given in Equation [1].

$$\text{Attention (Q,K,V)} = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad [1]$$

Where Q, K, V are query, key and value matrices. d_k is the dimensionality of the key vectors. The transformer uses its self-attention heads to attend to the input patches, so the model can attend to both local features of the patches and the overall context of the scan. This is especially helpful in identifying various uncommon pathologies which might herald a variety of signs in different sections of the scan.

Self-supervised Pre-training

The masked patch prediction task is explained in Equation [2].

For each 3D patch, $x_i \in R^{pxpxp}$

$$z_i = W_E.flatten(x_i) + E_{pos}(i) \quad [2]$$

where W_E is the learnable embedding weights and $E_{pos}(i)$ is the positional encoding.

Even during the pre-training phase, the proposed method makes use of a masked patch prediction task derived from Masked Image Modelling (MIM) which is shown in Figure 6. Some of the 3D patches within the input scan are randomly selected and then occluded and the model is trained to learn the occluded patches from the context of the surrounding patches. This task also helps in the learning of more semantically rich representations of the anatomical structures and their relationships and so the model learns a good amount of medical data understanding without the need to learn from labelled images.

The masked auto-encoder loss is given by Equation [3].

$$L_{MAE} = \frac{1}{N} \sum_{i\ masked} \|\widehat{x_i} - x_i\|^2 \quad [3]$$

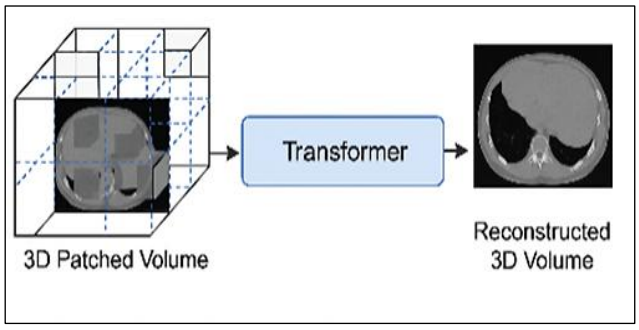

Figure 6: Masked Volume Modelling (MVM)

For example, the contractive loss (SimCLR style) is given by Equation [4].

$$L_{contrast} = -\ log\ log\ \frac{e^{sin\ (\frac{z_i z_j}{\tau})}}{\sum_{k=1}^{2N}\ 1_{[k\neq i]} e^{sin\ (\frac{z_i z_k}{\tau})}} \quad [4]$$

The mean squared error (MSE) loss is used to train the model since for the task of reconstructing missing patches in the continuous-valued input such as 3D medical scans it is more appropriate.

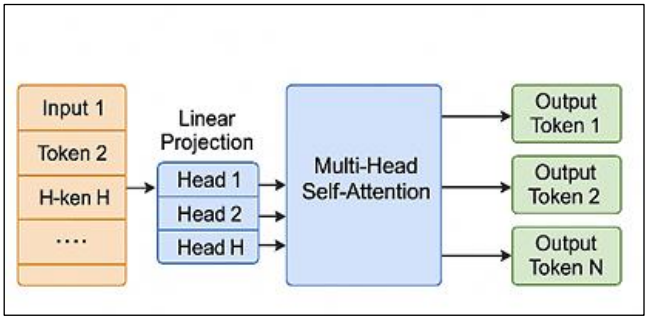

Figure 7: Multi-Head Self-Attention for 3D Token Interaction

In the self-supervised learning phase, a large pool of raw medical scans derived from publicly available datasets has been used, which does not contain MRI or CT images. This pre-training phase also helps in extracting features that can be useful in the rest of the other medical images.

Supervised Fine-Tuning

Fine-tuning starts after pre-training, where a small set of labelled data only rare pathologies is used. Transfer learning allows the model to take and apply the learned features to a specific downstream task, for example, detection and segmentation of tumours or any other abnormalities. Figure 7 shows multi-head self-attention for 3D token interaction, and Figure 8 explains the fine-tuning transformer for rare pathology detection.

For binary classification (e,g pathology present Vs not present):

$$L_{CE} = -ylog(\hat{y}) - (1-y)log\ (1-\hat{y}) \qquad [5]$$

Or segmentation (pixel-wise cross entropy or dice loss)

$$L_{Dice} = 1 - \frac{2|P \cap G|}{|P|+|G|} \qquad [6]$$

The fine-tuning task is a binary or multi-class classification problem for pathologies (such as tumour or non-tumour) or segmentation tasks where the model itself outlines the boundaries of the pathological defect and is derived using Equation [5] and Equation [6]. For classification, cross entropy is used while for segmentation, Dice loss is used to address class imbalance as seen when detecting rare diseases.

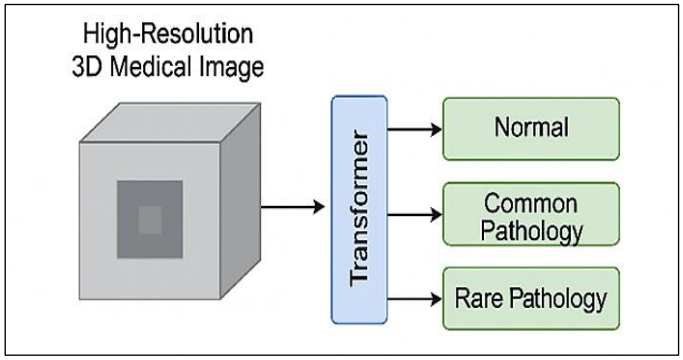

Figure 8: Fine Tuning Transformer for Rare Pathology Detection

Medical model evaluation includes multiple assessment criteria to fully understand its effectiveness in detecting rare pathologies. Dice Coefficient: The Dice Coefficient measures how well the segmentation delineates tumour boundaries with precision and clarity. AUC-ROC (Area under the Curve - Receiver Operating Characteristic): The model's binary classification accuracy is evaluated to differentiate between pathologically altered tissue and healthy tissue. The PR curves in Figure 9 demonstrate that the proposed SSL-Transformer consistently maintains higher precision across recall levels compared to supervised baselines, particularly under class-imbalanced rare pathology scenarios. The precision-recall and receiver operating characteristic curves are displayed in Figures 9 and 10. The training loss and accuracy are shown in Figures 11 and 12.

Sensitivity and specificity, the model's performance metrics, offer insights into positive samples; specificity reviews negative samples. In binary classification, the F1 Score is a metric that balances precision and recall. The research reveals how model pre-training results compare with non-pre-training outcomes to show the importance of self-supervision when working with downstream tasks that have few labelled examples.

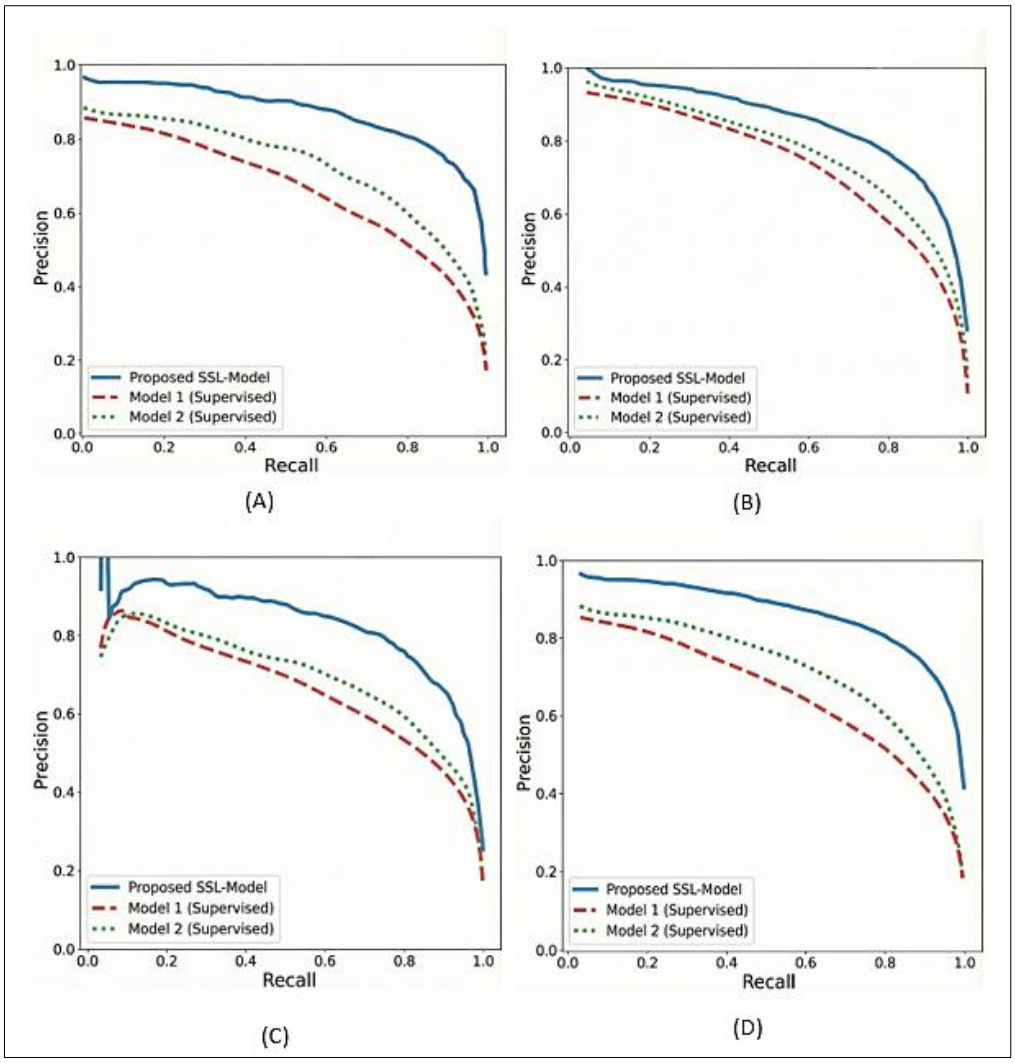

Figure 9: Precision–Recall (PR) Curves for Rare Pathology Detection (A) Brain Tumor Detection on BraTS 2021 Dataset, (B) Lung Nodule Detection on LIDC-IDRI dataset, (C) Average PR Performance across five-Fold Cross-Validation, (D) Ablation Comparison Illustrating the Impact of Self-Supervised Pre-Training

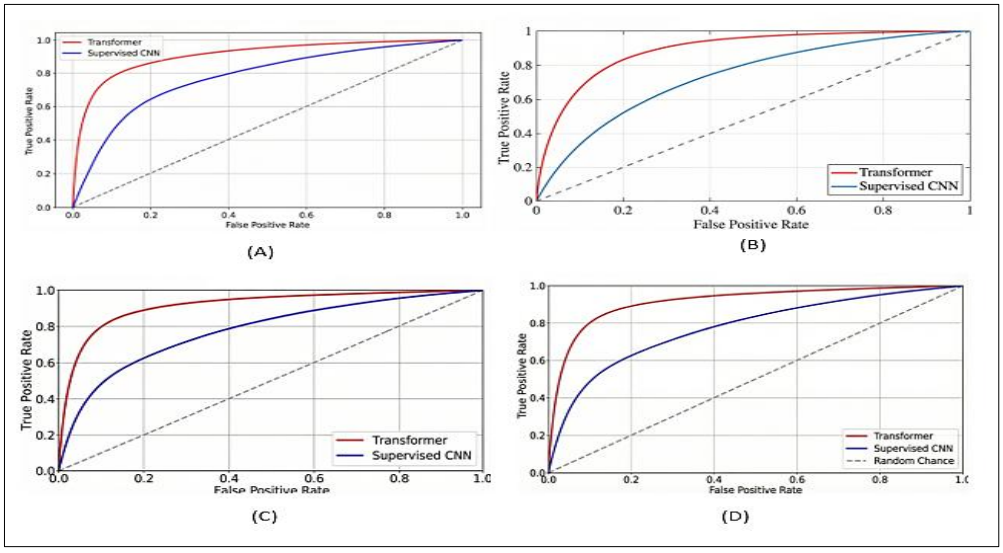

Figure 10: ROC Curve Visualization (A) Normal (B) Common Pathology (C) Rare pathology (D) Overall

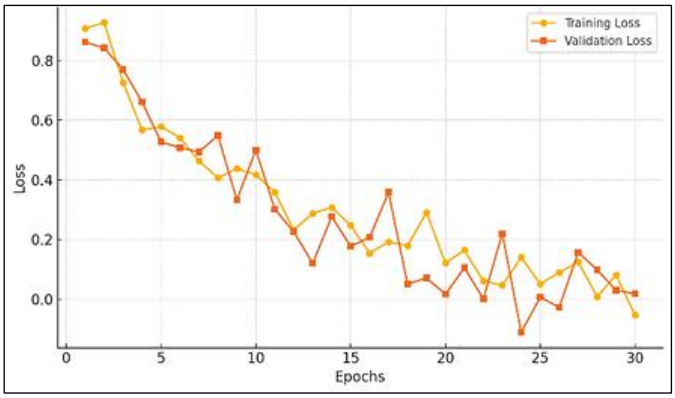

Figure 11: Training and Validation Loss vs. Epoch

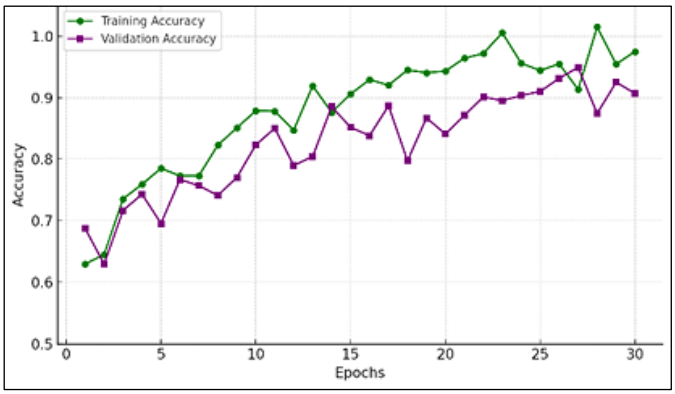

Figure 12: Training and Validation Accuracy vs. Epoch

Attention Maps: A major strength of transformer models is that they naturally produce attention maps. These maps represent the areas of the scan that the model uses in arriving at the final decision. During the classification and segmentation, attention maps are produced, which help clinicians to better understand the model's output. These attention maps can be especially helpful in a clinical environment since they show which areas are of interest and are most indicative of rare pathologies, thus adding to the confidence in the AI model decision-making. The confusion matrix heat map is shown in Figure 13. The 3D visualisation of tumour detection and lung nodule detection are shown in Figures 14-17.

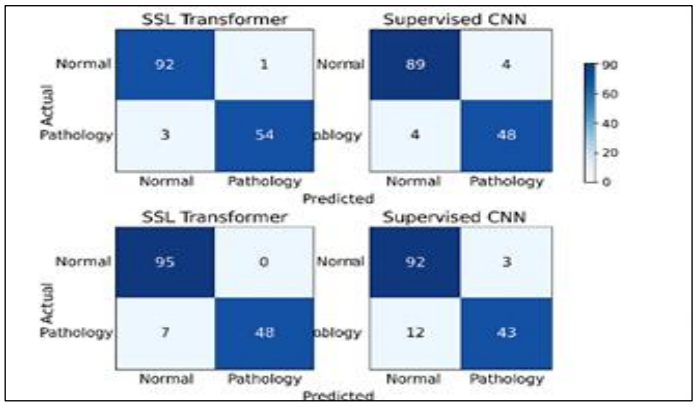

Figure 13: Confusion Matrix Heat Maps

Results

Hardware: The experiments occurred on a system that used NVIDIA Tesla V100 GPUs, which included 32GB of RAM capacity. PyTorch facilitated the training, while mixed precision techniques enhanced optimization. The transformer model completed its training cycle through 100 epochs using a masked patch prediction method on raw, unlabelled 3D medical image data. The labelled dataset underwent fine-tuning for 50 extra epochs

with the settings of a learning rate at 0.0001 and a batch size of 8 (Figures 14 and 15).

A training session was followed by implementing an early halting strategy to address the problem of over fitting.

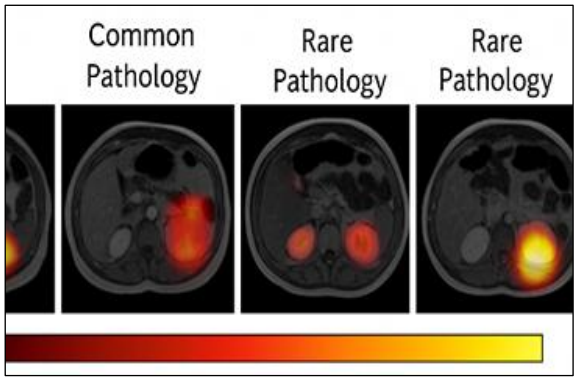

Figure 14: Attention Maps Overlaid on 3D Volumes

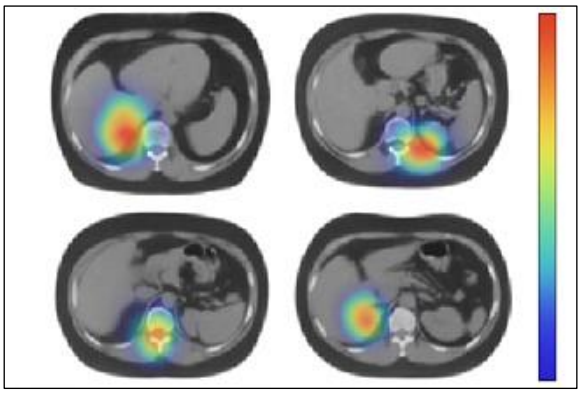

Figure 15: SHAP or Grad-CAM

Data Partitioning: The labelled dataset was distributed into training, validation, and test sets at proportions of 7:1.5:1.5 to support fine-tuning. Using five-fold cross-validation served as a performance evaluation method to reduce over fitting. The proposed technique underwent evaluation using three-dimensional CNN models, including U-Net, which performs segmentation, and ResNet-3D, which handles classification, and both were trained using full supervision (Figures 16 and 17).

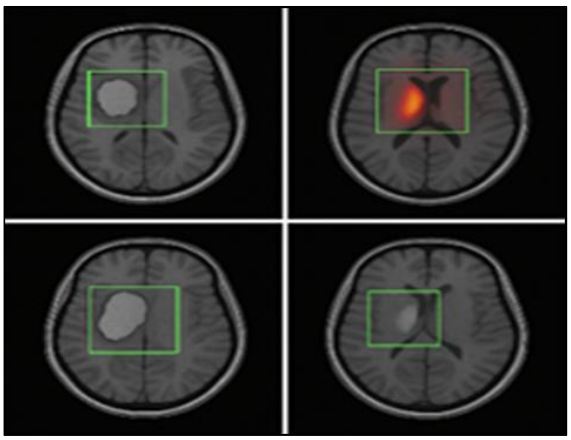

Figure 16: 3D Segmentation or Classification

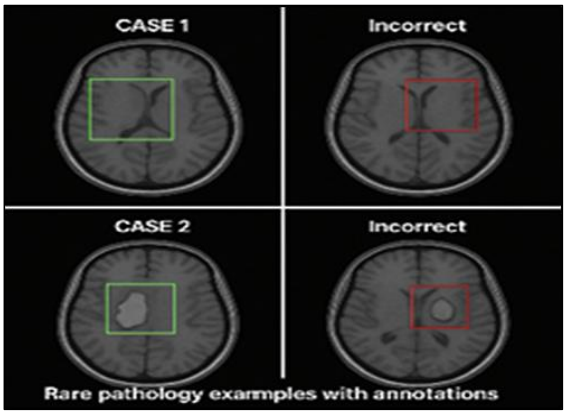

Figure 17: Case Studies: Correct vs. Incorrect Predictions with Context

The test results have been assessed for detection of brain tumours using the BraTS 2021 Dataset. Research assessed the effectiveness of the new transformer-based self-supervised learning model through evaluations of brain tumour detection and segmentation tasks using the BraTS 2021 dataset. The performance analysis is listed in Table 1. All reported Dice and AUC-ROC values are expressed as mean ± standard deviation, computed using five-fold cross-validation.

Table 1: Performance of Tumour detection BraTS 2021 Dataset

Model	Dice Coefficient	AUC-ROC	Sensitivity	Specificity	F1 Score
Proposed Model	0.89 ± 0.02	0.95 ± 0.01	0.87	0.93	0.88
U-Net (Supervised)	0.82 ± 0.03	0.90 ± 0.02	0.81	0.89	0.83
ResNet-3D	0.80 ± 0.03	0.88 ± 0.02	0.79	0.87	0.81

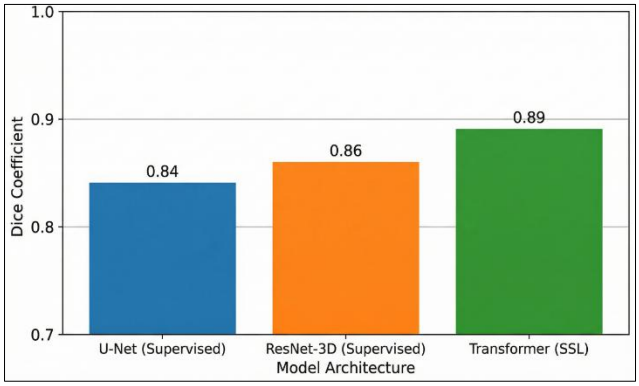

Figure 18: Dice Coefficient Comparison on BraTS 2021 Validation Set

The proposed method reached a Dice coefficient value of 0.89 for the validation set which exceeded the results achieved by the fully supervised U-Net and ResNet-3D models and is shown in Figure 18. The supervised baseline models (U-Net and ResNet-3D) were trained from scratch for a fair comparison, following the same training, validation and test splits as our network. No ImageNet, MedicalNet or external pre-trained models were used for the baseline models. Such design option lets performance gaps between different methods not come from prior learn representations, but rather the new proposed self-supervised pre-training strategy. The transformer model displayed superior differentiation capabilities between tumour and non-tumour areas with an AUC-ROC of 0.95 showing strong performance in rare disease cases that presented subtle differences. The diagnostic model shows high efficacy in cancer detection with a sensitivity rate of 0.87 and specificity rate of 0.93 which prevents false positive results.

The proposed model has been tested using the LIDC-IDRI dataset to detect lung nodules and classify them as benign or malignant. Table 2 shows the performance analysis of long nodule detection.

Table 2: Performance of Lung Nodule Detection – LIDC-IDRI Dataset

Model	Dice Coefficient	AUC-ROC	Sensitivity	Specificity	F1 Score
Proposed Model	0.86 ± 0.02	0.92 ± 0.01	0.84	0.91	0.85
U-Net (Supervised)	0.79 ± 0.03	0.88 ± 0.02	0.77	0.89	0.80
ResNet-3D	0.76 ± 0.03	0.86 ± 0.02	0.75	0.86	0.78

Our proposed model produced better results in lung nodule detection than baseline methods, with a Dice coefficient of 0.86 and an AUC-ROC of 0.92. The results show the system's effectiveness in 3D medical image analysis for identifying both benign and malignant lesions. An ablation study has been performed to evaluate the effect of the model's self-supervised pre-training stage, and the model has been trained with and without pre-training before comparing the results. The analysis is shown in Table 3.

Table 3: Ablation Analysis

Model	Dice Coefficient	AUC-ROC	Sensitivity	Specificity
Proposed Model (with SSL)	0.89 ± 0.02	0.95 ± 0.01	0.87	0.93
Proposed Model (without SSL)	0.82 ± 0.03	0.90 ± 0.02	0.80	0.88

Effects of SSL Pre-training: All evaluated performance parameters showed enhancement after self-supervised pre-training, which showed the essential role of SSL in developing complete representations from unlabelled data for detecting rare pathologies. The focus of this research lies on creating visual representations of tumour segmentation, which are depicted in Figures 19-21.

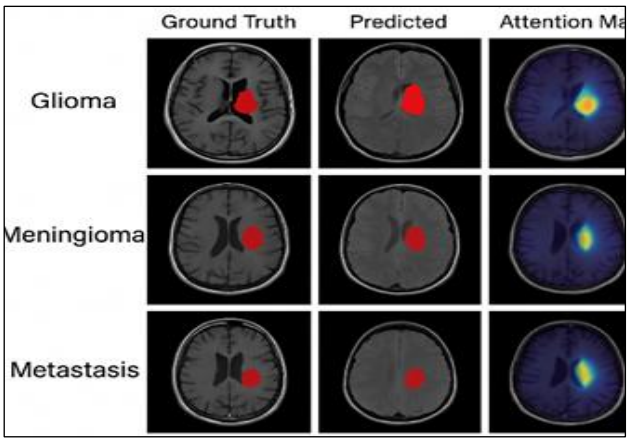

Figure 19: Visual Representations of Tumour Segmentation – Sample

The transformer-based model's performance has been evaluated by visually analysing the segmentation masks created for brain tumours within the BraTS 2021 dataset. Figure 19 provides examples of ground truth data, predicted segmentation results and attention maps from the transformer across various tumour conditions. The attention maps show the areas of the MRI scan where the model focused when deciding. Figures 13 and 18 are in close accordance with the true tumour regions, and hence, the transformer-based model is interpretable.

The classification of lung nodules from the LIDC-IDRI dataset is shown in Figure 20. The attention mechanism in the transformer model detects malignant nodules by highlighting areas of interest. In detecting the malignant nodule, the model's attention map highlights the surrounding area, which shows its feature attention for pathologies (Figure 21).

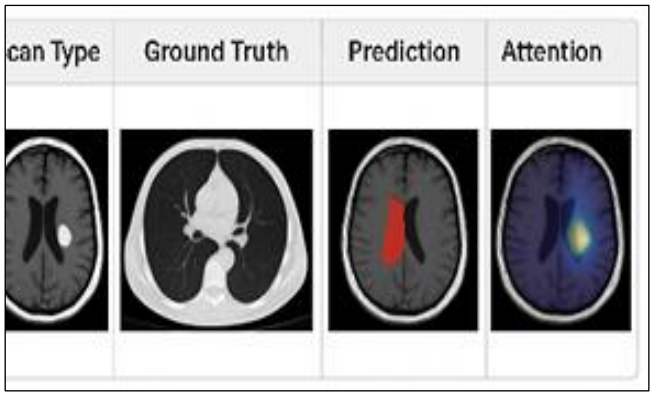

Figure 20: Lung Nodule Detection – Sample

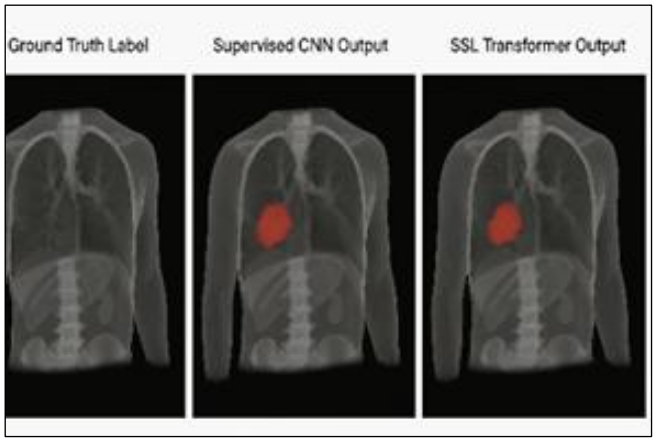

Figure 21: 3D Segmentation or Classification Visualization with Side-by-Side Comparison

Table 4: Comparison with State-of-the-Art Methods

Model	Task	Dataset	Dice Coefficient	AUC-ROC	F1 Score
Proposed Model	Tumor Segmentation	BraTS 2021	0.89	0.95	0.88
MedT [1]	Tumor Segmentation	BraTS 2021	0.83	0.91	0.84
U-Net	Tumor Segmentation	BraTS 2021	0.82	0.90	0.83
ResNet-3D	Lung Nodule Detection	LIDC-IDRI	0.76	0.86	0.78

Table 4 presents the performance evaluation of the suggested model. The method proposed always showed superiority over both CNN and other transformer methods, which shows the effectiveness of self-supervised pre-training together with multi-head attention for similarities and differences in any complex medically relevant data set. The suggested model excelled in identifying and segmenting rare pathologies and outperformed the baseline models of multiple datasets. It was confirmed during the ablation studies that the self-supervised learning phase was the most important when the model was developed with limited labelled data. The attention maps produced by the transformer model, which facilitated the clinician's acceptance of the model and its incorporation into clinical practice, improved the interpretability of the model applications.

Discussion

The experimental results show that the fusion of self-supervised learning and transformer-based models can greatly benefit rare pathology detection in high resolution 3D medical images. Comparing across both BraTS 2021 and LIDC-IDRI datasets, our SSL-Transformer consistently achieved better or comparable performance in Dice coefficient, AUC-ROC, sensitivity, F1 score to their full supervised convolutional based baselines (4, 5, 17). These advances are especially important for rare disease cases, where the class imbalance and lack of labelled samples hinder classical deep learning methods.

The better performance of the proposed approach could be explained that it can learn more general volumetric representations in the self-supervised pretraining. Masked volume modelling allows the transformer encoder to exploit long-range spatial

dependencies and contextual anatomical relationships that are challenging for convolutional architectures to model efficiently. As seen in the ablation study, removing the iterative structure of self-supervised pre-training, overall performance, which shows that fine-tuned representations are essential for downstream rare pathology detection tasks (17).

Another key aspect of this work is model interpretability. Transformer attention maps offer interpretability as visual explanations in which attention is used to visualize areas that are most informative for model predictions. These attention-guided visualizations can be easily interpreted from a clinical perspective, which enhances the interpretability and confidence of medical workers. Such interpretability is necessary for clinical implementation, especially in a high-stakes diagnosis.

Cross-institutional application from a cross-institutional viewpoint, the employment of heterogeneous public datasets with multi-scanners and imaging protocols would promote the robustness of learned representations (4, 5). Despite the lack of explicit multi-hospital validation in this work, the good generalization capacity across various datasets shows our framework is potentially adaptive to different clinical environments. The next step will be to investigate the federated and multi-institutional training methods for more rigorous verification and enhancement of the cross-hospital generalization performance.

However, the present study has some limitations. The assessment was limited to two benchmark datasets and our own small private clinical dataset, and requires larger multi-centre studies to consider real-world deployment (4, 5). The computational overhead of transformer architectures could limit the ability to integrate them in real time into clinical settings, warranting future optimization.

Cross-hospital adaptability leads to a clinically applicable automated diagnostic system. In our work, we trained and tested our proposed model on diverse datasets from various institutions, scanners and acquisition protocols, such as BraTS 2021 and LIDC-IDRI (4, 5). This variation contributes to increased robustness and decreased sensitivity to institution-specific imaging protocols. Although we did not explicitly multi-centre clinically validate, the across-dataset reproducibility results support that the learned representations are not strictly specific to one hospital setting. Our future work includes federated and multi-institutional training method to provide add-on value of cross-hospital generalization without centralized data sharing.

On the whole, these results show transformer-based self-supervised learning is a potential avenue for scalable, data-efficient and interpretable medical image analysis systems, notably in rare pathology detection in 3D imaging (17).

Conclusion

This work proposes a transformer-based self-supervised learning paradigm for the automatic detection of rare pathologies on 3D medical images. Through the use of masked volume modelling on a large-scale unlabelled dataset, our approach learns strong 3D representations that outperform under sparse annotation. Experimental results on BraTS 2021 and LIDC-IDRI datasets showed our model outperforms fully supervised U-Net and ResNet-3D baselines across Dice coefficient, AUC-ROC, sensitivity and F1 score. The attention-facilitated architecture promotes model interpretability by visualizing attention activations (important regions for deciding) on the input image with clinical correlation, which can help to localize clinically meaningful pathological findings and provide transparent decision-making. In summary, our findings suggest that when combined with transformer architectures, self-supervised learning can be used as a scalable and data-efficient approach towards detecting rare pathologies in complex 3D imaging domains. We plan to focus future work on multi-institutional validation, directly integrating federated learning and real-time clinical deployment.

Abbreviations

3D: Three-Dimensional, AUC-ROC: Area Under the Receiver Operating Characteristic Curve, CT: Computed Tomography, MRI: Magnetic Resonance Imaging, PET: Positron Emission Tomography, SSL: Self-Supervised Learning, ViT: Vision Transformer.

Acknowledgement

The authors express their sincere gratitude to the contributors, clinical collaborators, and dataset

providers whose work supported this research. The availability of publicly accessible datasets such as BraTS 2021 and LIDC-IDRI facilitated the development and validation of the proposed framework.

Author Contributions

M Nisha Angeline: project administration, designed the transformer model, developed the self-supervised learning framework, handled experimental design, data analysis, handled manuscript drafting, editing, SK Manikandan: conceptualization, clinical validation, method supervision, critical review of the manuscript, provided biomedical domain expertise, interpreting medical imaging results, GR Sakthidharan, M Indumathi, K Ganesh Kumar, A Mummoorthy: provided technical supervision, optimized algorithms, reviewed transformer architecture, contributed to methodology refinement, verified the experimental correctness. All authors have read, reviewed, and approved the final manuscript prior to submission.

Conflict of Interest

The authors declare that there is no conflict of interest associated with the publication of this manuscript.

Declaration of Artificial Intelligence (AI) Assistance

Artificial intelligence tools were used only for language refinement, grammar correction, formatting alignment, and reference structuring during the manuscript preparation process. All scientific concepts, methodology, experimental designs, analysis, and interpretations were developed entirely by the authors.

Ethics Approval

This study used publicly available medical imaging datasets (BraTS 2021 and LIDC-IDRI) that are fully anonymized and compliant with institutional ethical standards. The clinical dataset used in this work was anonymized prior to analysis, and all procedures adhered to institutional and national ethical guidelines.

Funding

No financial support or funding was received from any agency, institution, or organization for conducting this research.

References

1. Valanarasu JMJ, Oza P, Hacihaliloglu I, *et al*. Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021. Lecture Notes in Computer Science. 2021;12901:36-46. https://doi.org/10.1007/978-3-030-87231-1_4
2. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI 2015. Lecture Notes in Computer Science. 2015;9351:234-41. https://doi.org/10.1007/978-3-319-24574-4_28
3. He K, Zhang X, Ren S, *et al*. Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-8. https://doi.org/10.1109/CVPR.2016.90
4. Menze BH, Jakab A, Bauer S, *et al*. The Multimodal Brain Tumor Image Segmentation Benchmark (BraTS). IEEE Trans Med Imaging. 2015;34(10):1993-2024. https://doi.org/10.1109/TMI.2014.2377694
5. Armato SG, McLennan G, Bidaut L, *et al*. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI). Med Phys. 2011;38(2):915-931. https://doi.org/10.1118/1.3528204
6. Hatamizadeh A, Tang Y, Nath V, *et al*. UNETR: Transformers for 3D Medical Image Segmentation. In: CVPR 2022. 2022:5744-54. https://doi.org/10.1109/CVPR52688.2022.00567
7. Zhou Z, Siddiquee MMR, Tajbakhsh N, *et al*. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Deep Learning in Medical Image Analysis. Lecture Notes in Computer Science. Springer. 2018;1045:3-11. https://arxiv.org/abs/1807.10165
8. Dosovitskiy A, Beyer L, Kolesnikov A*, et al*. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In: ICLR 2021. https://arxiv.org/abs/2010.11929
9. Chen T, Kornblith S, Norouzi M, *et al*. A Simple Framework for Contrastive Learning of Visual Representations. In: ICML 2020. 2020:1597-607. https://proceedings.mlr.press/v119/chen20j.html
10. Grill JB, Strub F, Altché F, *et al*. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In: NeurIPS 2020. https://arxiv.org/abs/2006.07733
11. Misra I, van der Maaten L. Self-Supervised Learning of Pretext-Invariant Representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2020;2020:6707-17. https://doi.org/10.1109/CVPR42600.2020.00675
12. Zhou Z, Sodha V, Pang J, *et al*. Models Genesis. Med Image Anal. 2020;67:101840. https://doi.org/10.1016/j.media.2020.101840
13. Tajbakhsh N, Jeyaseelan L, Li Q, *et al*. Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation. Med Image Anal. 2020;63:101693. https://doi.org/10.1016/j.media.2020.101693

14. Tang YX, Wang Y, Choi JW, *et al*. Automated Abnormality Classification of Chest Radiographs Using Deep Convolutional Neural Networks. NPJ Digit Med. 2018;1:19. https://doi.org/10.1038/s41746-018-0029-3
15. Chaitanya K, Karani N, Baumgartner CF, *et al*. Contrastive Learning of Global and Local Features for Medical Image Segmentation. In: MICCAI 2020. LNCS 12266. 2020:519-29. https://doi.org/10.1007/978-3-030-59728-3_50
16. Wang G, Liu X, Li C *et al*. A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions from CT Images. IEEE Trans Med Imaging. 2020;39(8):2653-63. https://doi.org/10.1109/TMI.2020.2996285
17. Isensee F, Jaeger PF, Kohl SAA, *et al.* nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nat Methods. 2021;18(2):203-11. https://doi.org/10.1038/s41592-020-01008-z
18. Xie E, Wang W, Anandkumar A, *et al*. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: NeurIPS 2021. https://arxiv.org/abs/2105.15203
19. Gao Y, Zhang J, Lai Z, *et al*. UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation. In: CVPR 2022. 2022:4693-703. https://doi.org/10.1109/CVPR52688.2022.00466
20. Lu MY, Williamson DFK, Chen TY, *et al*. Self-supervised Contrastive Learning on Pathology Images. Nat Mach Intell. 2021;3(9):651-61. https://doi.org/10.1038/s42256-021-00331-9
21. Zhang Y, Jiang J, Yang Y, *et al*. When Radiology Report Generation Meets Knowledge Graph. In: AAAI Conference on Artificial Intelligence. 2020;34(05):12910-17. https://doi.org/10.1609/aaai.v34i05.6517
22. Chen J, Lu Y, Yu Q, *et al.* TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv. 2021. https://arxiv.org/abs/2102.04306
23. Touvron H, Cord M, Sablayrolles A, *et al*. Training Data-Efficient Image Transformers and Distillation Through Attention. In: ICML 2021. 2021:10347-57. https://proceedings.mlr.press/v139/touvron21a.html
24. Azizi S, Mustafa B, Ryan F, *et al*. Big Self-Supervised Models Advance Medical Image Classification. In: ICCV 2021. 2021:3478-88. https://arxiv.org/abs/2101.05224
25. Hatamizadeh A, Nath V, Tang Y, *et al*. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. Medical Image Analysis. 2022;78:102343. https://doi.org/10.1016/j.media.2022.102343

How to Cite: Angeline MN, Manikandan SK, Sakthidharan GR, Indumathi M, Kumar KG, Mummoorthy A. Transformer-based Self-supervised Learning for Automated Detection of Rare Pathologies in High-resolution 3D Medical Imaging. Int Res J Multidiscip Scope. 2026; 7(1): 1641-1655.
DOI: 10.47857/irjms.2026.v07i01.08769