3DViTMedNet: Hybridizing 3D CNN with Vision Transformer for Advanced Medical Image Classification

Spencer Gerontzos and Son Tran
School of Information Technology, Deakin University

***Abstract*—In this project, we present 3DViTMedNet, a novel three-dimensional medical image classification approach that integrates both 3D Convolutional Neural Networks (CNN) and Transformer architectures. This hybrid model is designed to effectively process 3D medical images by leveraging the strengths of each architecture. The 3D CNN component captures 3D representations that are inherent to the modality of the input data, enabling the model to extract vital spatial features across different anatomical planes—coronal, sagittal, and axial. These 2D slices are then processed using pretrained 2D models to harness the local relationships present within each plane. The slices extracted from the 3D CNN are embedded using a 2D CNN, after which they are tokenized and passed through a Vision Transformer, allowing the model to capture both global and local relationships within the data. This architectural design ensures that 3DViTMedNet retains the 3D representational features that might otherwise be lost in purely 2D models, while also capitalizing on the global attention mechanisms of the Transformer to model intricate dependencies across the image. The combination of the 3D CNN for local context and the Vision Transformer for global context facilitates a more comprehensive analysis of the medical images. Empirical studies demonstrate that 3DViTMedNet can be competitive with and outperform traditional CNN-based architectures, such as ResNet, by leveraging both 3D spatial awareness and attention mechanisms. This highlights the model's ability to effectively handle the complexities of 3D medical imaging data, making it a promising approach for future advancements in medical image classification.**

***Index Terms*—Vision Transformer, Convolutional Neural Network, ResNet, Attention, 3DViTMedNet.**

I. Introduction

Convolutional Neural Networks (CNNs) have established themselves as leading performers for Computer Vision (CV) applications, demonstrating exceptional capabilities in tasks such as segmentation, classification, and object detection. Recently, CNN-based architectures have gained significant traction in medical imaging, allowing researchers to efficiently analyze a diverse array of imaging modalities, including CT, MRI, and PET scans, in various dimensions. Numerous approaches have been explored within the medical field for analyzing 3D images using both 2D and 3D methodologies. Architectures employing 2D techniques often leverage pre-trained models that have been trained on large datasets, typically from different domains. However, these architectures face challenges in effectively capturing the full spatial context of 3D images. While 3D methods are capable of providing detailed 3D representations, preserving natural local relationships, there is a notable scarcity of pre-trained 3D models that can be utilized to enhance model performance as well as a significant increase in computational cost.

Self-attention-based transformers have demonstrated success in natural language processing tasks. Vision Transformers (ViTs) have similarly gained popularity in various CV tasks, offering an increased receptive field compared to traditional CNN models. ViTs have demonstrated the capability to achieve results comparable to widely used models. Limitations in pure ViT architectures arise due to a lacking inductive bias, necessitating large training samples (and resources) for effective optimization. Additionally, these models often struggle with incorporating local features within each patch, degenerating their performance in comparison to CNN counterparts.

Given the limited size and availability of 3D medical image datasets, Convolutional Neural Network (CNN) models typically demonstrate superior performance metrics. This observation suggests that pure Vision Transformer (ViT) models face challenges in learning meaningful representations from such data, primarily due to the previously mentioned limitations.

Hybrid architectures that combine the strengths of CNNs and ViTs have been proposed to address these challenges. CNNs, with their strong inductive biases and locality, can achieve high performance even with limited data. However, their relatively small receptive field restricts their ability to capture long-range dependencies within the data. In contrast, Transformers have minimal inductive bias, which can hinder performance on smaller datasets, but their self-attention mechanism enables them to consider a larger input area, capturing more general patterns and relationships with a high receptive field. Most of these hybrid architectures employ 2D CNN blocks followed by a transformer model. In contrast, our model integrates both 3D CNN and 2D CNN components to capture 3D and local representative features, respectively. This design allows for the generation of a rich encoder prior to the transformer block, enhancing the model's ability to process 3D medical images effectively.

The analysis of biomedical images has long presented challenges, primarily due to the need for specialized knowledge and training to interpret these images accurately. Even

with this expertise, various imaging modalities exhibit subtle variations that can indicate different medical conditions. For instance, Figure 2 illustrates a comparison between various rib fractures, emphasizing the importance of recognizing the subtleties of the respective fractures.

Transformer models are well-suited to map long-range dependencies, enabling a comprehensive understanding of the overall structure within the dataset. Complementing this, CNN backbones excel at detecting local features. In the context of Figure 2, the transformer can capture the broader anatomical structure, such as the spacing between the ribs and their alignment. Meanwhile, the CNN can focus on identifying specific local anomalies, such as fractures, which may occur in isolated regions. This combination allows for a more nuanced analysis of complex medical images.

In this study, we introduce 3DViTMedNet, a medical image classifier designed for 3D imaging modalities using the MedMNIST3D database. The contribution of our research is twofold:

- We propose 3DViTMedNet tosuccessfully integrate both CNN and transformer architectures, leveraging the strength of vision transformers in capturing global relationships and the high inductive bias of CNNs for local feature extraction. This synergy leads to superior performance when compared to pure transformer-based models.
- We provide interpretability by visualizing 3DViTMedNet's attention maps and comparing them to those of benchmark models. The alignment of attention with key regions of interest enhances the understanding of the model's focus, providing insights into its decision-making process and highlighting critical areas for classification.

II. Related Works

A. CNN in Medical Image Classification

In recent years, deep learning models, particularly CNNs, have achieved remarkable success across various medical imaging tasks, including segmentation, detection, registration, and classification. These models excel at uncovering latent representations and effectively identifying disease-related pathologies, making them powerful tools for medical image analysis.

Nie et al. developed a CNN encoder for feature extraction, which was subsequently used to train a support vector machine (SVM) for predicting survival outcomes in patients with high-grade gliomas [1]. Similarly, Wegmayr et al. proposed an InceptionNet-inspired architecture [2] employing varying kernel sizes to classify patients with Alzheimer's disease [3]. Islam and Zhang introduced an ensemble of three CNN models, each designed to classify images from the coronal, sagittal, and axial planes, with a majority voting scheme determining the final classification for Alzheimer's disease diagnosis [4]. Kruthika et al. proposed an ensemble approach combining Capsule Networks, CNNs, and pretrained autoencoders to classify Alzheimer's disease [5]. Gao et al. applied convolutional layers to 2D and 3D medical images of Alzheimer's patients, fusing the outputs from both approaches to produce a final classification [6].

A notable limitation of CNN-based methods is their tendency to overlook global relationships due to the inherently local nature of convolutional operations. Korlev et al. demonstrated that ResNet [7] and VGGNet [8] models, when used for Alzheimer's disease classification, focus primarily on localized regions, such as the hippocampus, while failing to capture global structural changes like the shrinking of the cerebral cortex [9]. Similarly, Yang et al. employed ResNet and VGGNet for Alzheimer's classification and reached comparable conclusions regarding the models' inability to recognize broader anatomical relationships [10].

B. Approaches to 3D Medical Image Representation

2D CNN models have been extensively utilized for 3D image classification tasks, largely due to the availability of pretrained models. In recent years, various representation methods have been developed to leverage the power of these pretrained models. One common approach involves iteratively extracting slices from one of the three dimensions (axial, coronal, sagittal), which has demonstrated effectiveness across numerous domains. However, this method inherently results in a loss of 3D context and spatial information due to the isolated nature of the slices [11], [12].

To address the lack of 3D context, another widely adopted method involves collecting slices from each anatomical plane (axial, coronal, sagittal) and stacking these slices as individual channels for processing by 2D CNN models [13]. While this approach allows the model to receive information from multiple key anatomical views, the slices across planes are not spatially aligned, which can hinder the model's ability to develop true spatial relationships between features from the coronal, sagittal, and axial perspectives.

Another popular method is the multi-slice representation, where multiple consecutive slices are extracted from a single plane and treated as separate input channels, similar to the RGB channels in traditional 2D images [14]. This approach ensures spatial alignment across slices, preserving the relationships within a specific plane. However, it limits the analysis to that single plane, potentially missing valuable information from the other anatomical views.

A critical limitation shared by these methods is that 2D CNNs are fundamentally unable to fully account for the native 3D structure of the input data. In contrast, 3D CNN models inherently learn 3D spatial features, making them well-suited for tasks such as medical image analysis, where accurate representation of 3D anatomical structures is critical for classification. However, despite their architectural advantages, the performance of 3D CNN models is often limited by the lack of large, annotated 3D datasets necessary for effective training as well as large parameters and computational costs.

C. Vanilla Transformers in Medical Image Classification

Following the remarkable success of transformer models in natural language processing (NLP) [15], substantial re-

search efforts have shifted towards adapting transformers for computer vision (CV) tasks. Pioneering studies, including ViT [16], SWIN [17] and DEiT [18] have been extensively explored in the domain of 3D medical image classification, demonstrating their potential in this field.

Gheflati et al [19]. employed a vanilla Vision Transformer (ViT) for classification of breast ultrasound images, demonstrating superior performance in terms of accuracy (ACC) and area under the curve (AUC) when compared to ResNet-based models. Similarly, Shome et al. applied a ViT to a custom COVID-19 dataset to differentiate between pneumonia, healthy, and COVID-19 cases, achieving results that outperformed popular CNN architectures [20].

These models, along with other 2D transformer-based approaches, benefit from the availability of large datasets, as transformers typically require significant amounts of data to surpass their CNN counterparts [21]. However, working with 3D datasets presents challenges due to the limited availability of annotated data. As a result, researchers often attempt to represent 3D images in a 2D format for ease of processing. Gao et al. demonstrated this approach by performing binary classification on COVID-19 patients, where slices from a single dimension were processed using a ViT, and a majority voting procedure was employed to determine the final classification [11]. Zhang and Wen [12] similarly used a slice-based method however with Swin Transformers for binary classification of COVID-19 patients, where 2D slices served as input to the transformer model, resulting in only marginal improvements over CNN-based models.

D. Hybrid Transformers in Medical Image Classification

Vanilla Vision Transformers struggle in medical image classification due to low inductive bias, which limits their ability to capture local spatial features [22]. Hybrid models, combining transformers with convolutional layers, improve local feature extraction while retaining global attention, making them more effective for medical tasks [23].

Dai et al. introduced TransMed, a ResNet-DeiT hybrid model designed to fuse slices from multiple modalities for genetic classification tasks [24]. Hsu et al. employed volumetric slices as input to a pretrained ResNet feature extractor, followed by classification using a Swin Transformer [25]. Jang and Hwang proposed a Cross-plane, multi-slice strategy, applying 3D convolutions to extract 3D representational features before extracting the slices [26]. Similarly, Jun et al. leveraged the multi-slice strategy, with both models utilizing CNNs for feature extraction prior to final classification via the Vision Transformer (ViT) [14].

To address the limitations of processing 3D images, we first extract 3D representational features using a 3D CNN, followed by a hybrid approach that combines multi-plane and multi-slice representations for more comprehensive image analysis. To overcome the lack of global contextual understanding inherent in CNNs, and the limited local context from transformers, we apply a transformer network to analyze and integrate the features extracted from these slices and planes, providing a more holistic representation.

III. Method

In this paper, we propose **3DViTMedNet**, as shown in1. Our model consists of five key modules designed to enhance performance in 3D medical image classification. First, a comprehensive data augmentation pipeline is implemented to improve the model's robustness and performance III-A. Second, image shifting techniques are applied to help the model capture more diverse and meaningful relationships in the data III-B. The third module is a 3D feature extractor that processes volumetric data, capturing critical 3D spatial features III-C. In the fourth module, the 3D input is divided into slices, which are tokenized and processed through a pretrained 2D CNN to extract 2D representational features III-D. In the fifth and final module, the tokens are passed into a standard ViT, which captures global relationships within the data, and the output of the transformer is processed through an MLP head to produce the final classification III-E.

A. Data Augmentation Pipeline

A data augmentation pipeline is crucial for enhancing model performance in 3D medical imaging by simulating real-world variability and reducing overfitting. In our approach, we apply transformations such as rotation, scaling, translation, noise addition, and flipping. Rotation is used to vary the orientation of the image (denoted as $\mathbf{X} \in \mathbb{R}^{C\times D\times H\times W}$), represented as:

$$\mathbf{X}_{\text{rotated}} = R(\theta_x, \theta_y, \theta_z) \cdot \mathbf{X}$$

where $R(\theta_x, \theta_y, \theta_z)$ represents rotation along the x, y, and z axes. Scaling adjusts image size, expressed as:

$$\mathbf{X}_{\text{scaled}}(x', y', z') = \mathbf{X}(sx, sy, sz)$$

Translation shifts the image along spatial axes, and noise simulates real-world imperfections:

$$\mathbf{X}_{\text{noisy}} = \mathbf{X} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise. Flipping mirrors the image across an axis. This augmentation pipeline increases model robustness by making it more adaptive to orientation, scale, position, noise, and symmetry variations found in 3D medical data.

B. Image Shifting

A shifted patch augmentation strategy to enhance the model's ability to capture spatial relationships within 3D medical images. Given an input 3D image $\mathbf{X} \in \mathbb{R}^{C\times D\times H\times W}$, where C is the number of channels, and D, H, and W represent the depth, height, and width of the image respectively, we generate shifted versions of the image along each axis. This is achieved by applying shifts to the depth, height, and width, denoted as Δ_d, Δ_h, and Δ_w, respectively. For instance, the image shifted along the depth axis is represented as $\mathbf{X}^{(D)}_{\text{shifted}}(d', h, w) = \mathbf{X}(d + \Delta_d, h, w)$, while shifts along

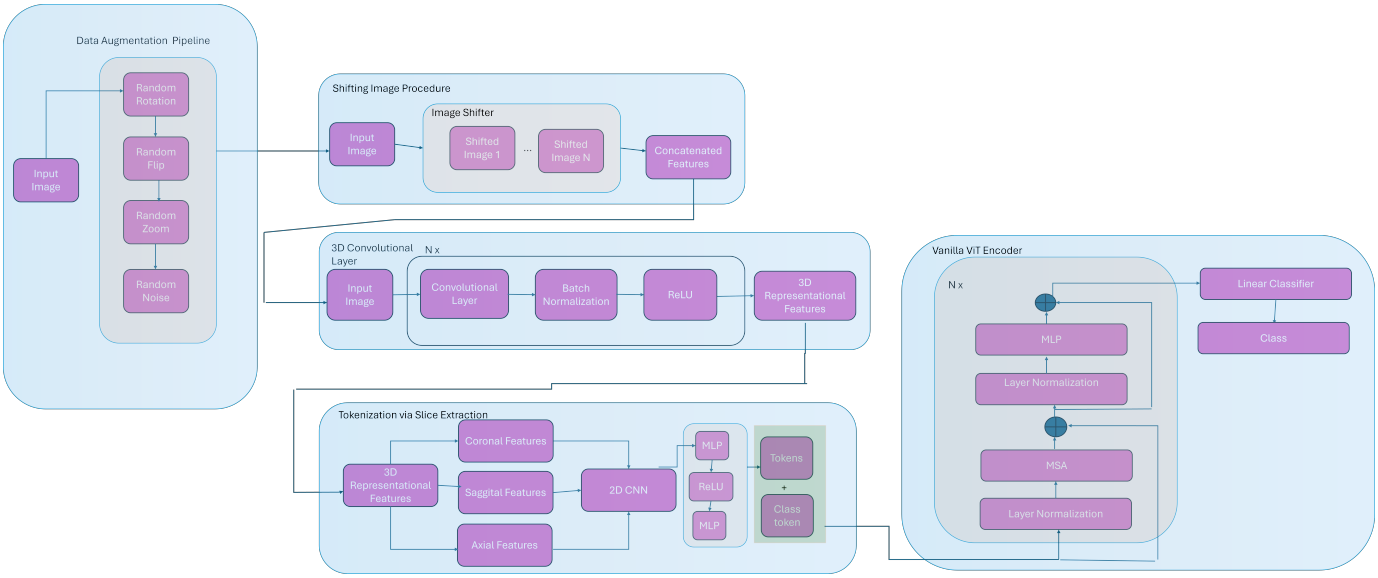

Fig. 1. Design of the 3DViTMedNet architecture

the height and width are expressed similarly as $\mathbf{X}^{(H)}_{\text{shifted}}$ and $\mathbf{X}^{(W)}_{\text{shifted}}$.

In addition to individual axis shifts, we apply combined shifts across multiple dimensions, such as along both depth and height $\mathbf{X}^{(D,H)}_{\text{shifted}}(d', h', w) = \mathbf{X}(d + \Delta_d, h + \Delta_h, w)$ and across all three dimensions $\mathbf{X}^{(D,H,W)}_{\text{shifted}}(d', h', w') = \mathbf{X}(d + \Delta_d, h + \Delta_h, w + \Delta_w)$.

The resulting augmented input tensor $\mathbf{X}_{\text{aug}} \in \mathbb{R}^{nC \times D \times H \times W}$ is formed by concatenating the original image and its shifted versions along the channel dimension. This allows the model to process the input from multiple perspectives, ultimately enhancing its capacity to learn richer spatial representations from the 3D medical images. The final augmented input is expressed as:

$$\mathbf{X}_{\text{aug}} = \text{Concat}\Big(\mathbf{X}, \mathbf{X}^{(D)}_{\text{shifted}}, \mathbf{X}^{(H)}_{\text{shifted}}, \mathbf{X}^{(W)}_{\text{shifted}}, \mathbf{X}^{(D,H)}_{\text{shifted}}, \mathbf{X}^{(H,W)}_{\text{shifted}}, \mathbf{X}^{(D,H,W)}_{\text{shifted}}\Big).$$

This augmentation improves the model's ability to capture fine-grained details in 3D medical images by leveraging multiple spatial contexts.

C. 3D Convolutional Layer

3D convolutional blocks are applied to extract feature representations from volumetric data by applying a convolution filter over the depth, height, and width of the input. Given an input tensor $X \in \mathbb{R}^{C \times D \times H \times W}$, where C is the number of channels, and D, H, and W represent the depth, height, and width, a 3D convolution applies a filter to produce a feature map $Y \in \mathbb{R}^{C' \times D' \times H' \times W'}$. In addition to this, after each 3D convolution, a ReLU activation function is applied to introduce non-linearity, and batch normalization is used to standardize the activation's, ensuring faster convergence and improved model stability.

D. Tokenization via Slice Extraction

After splitting the 3D input tensor $X \in \mathbb{R}^{C \times D \times H \times W}$ into 2D slices along the depth, height, and width planes via our slice extraction scheme, the resulting slices $S_d, S_h, S_w \in \mathbb{R}^{C \times H \times W}$ are passed into a pretrained 2D CNN encoder. This 2D CNN extracts localized feature representations from each slice using pretrained weights, yielding a feature map. The output feature maps are then processed through a non-linear projection via two multi-layer perceptron (MLP) blocks.

The non-linear projection is defined as:

$$F' = \text{ReLU}(W_1 F + b_1)$$

where W_1 is the weight matrix, b_1 is the bias term, and the ReLU activation introduces non-linearity into the projection. This is followed by another linear transformation:

$$\hat{F} = W_2 F' + b_2$$

where W_2 is the weight matrix and b_2 is the bias for the second projection. The result of these transformations is a tokenized output for each slice. These tokens T_d, T_h, T_w represent the feature maps from the depth, height, and width planes, respectively, and serve as compact, meaningful representations of the input.

E. Vision Transformer for Token Processing

The aforementioned tokens are sequentially passed through each transformer layer, consisting of multi-head self-attention, layer normalization, and MLP blocks, while maintaining consistent token dimensions across all layers. The final output is then processed through a standard MLP block to generate the final classification, where the number of output classes is determined by the specific dataset being utilized.

IV. Experiments

A. Datasets

The MedMNIST collection comprises six pre-processed datasets featuring a variety of imaging modalities, including CT, MRI, and electron microscopy. These datasets are designed for classification tasks, ranging from multi-class to binary classification. The dataset sizes vary between 1,200 and 2,000 samples. As illustrated in figure 3, the diversity of these datasets provides an excellent foundation for a wide range of classification challenges. The datasets have been pre-processed and split into training, validation, and test sets according to the methodology outlined in [27].

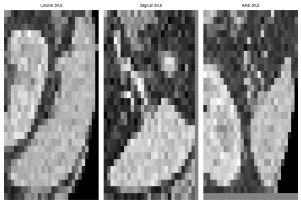

Fig. 2. OrganMNIST3D: Coronal, Sagittal, and Axial Plane Examples from the MedMNIST3D Dataset.

OrganMNIST3D is derived from the Liver Tumor Segmentation Benchmark (LiTS) [28]. This dataset is designed for multi-class classification across 11 body organs, comprising a total of 1,742 samples, shown in Figure 2.

NoduleMNIST3D is derived from the LIDC-IDRI31 dataset [29], a comprehensive public dataset of lung nodules captured through thoracic CT scans. It is designed for tasks such as lung nodule segmentation and binary classification of malignancy, based on a 5-level malignancy scale. The dataset includes a total of 1,633 samples.

AdrenalMNIST3D is a 3D shape classification dataset, comprising shape masks from 1,584 adrenal glands (792 patients), including both left and right glands. The data was collected from Zhongshan Hospital, and each 3D adrenal gland shape is annotated by an expert endocrinologist using abdominal CT scans. The dataset also includes binary classification labels, indicating whether the adrenal gland is normal or has an adrenal mass.

FractureMNIST3D is derived from the RibFrac Dataset [30], which includes approximately 5,000 rib fractures identified in 660 CT scans. The dataset classifies rib fractures into four clinical categories: buckle, nondisplaced, displaced, and segmental fractures.

VesselMNIST3D is based on the IntrA33 dataset [31], an open-access collection of 3D intracranial aneurysm models. This dataset comprises 103 3D brain vessel models, reconstructed from MRA images. A total of 1,694 healthy vessel segments and 215 aneurysm segments were automatically generated from these complete models.

SynapseMNIST3D is a 3D volume dataset based on the MitoEM dataset [32] designed to classify synapses as either excitatory or inhibitory. The dataset consists of 3D image volumes of an adult rat, obtained using a multi-beam scanning electron microscope, and includes a total of 1,759 samples.

B. Implementation details

In our implementation of the 3DViTMedNet architecture, key hyperparameters were selected to optimize performance. We used the Adam optimizer with a learning rate of 1×10^{-5} , combined with a MultiStepLR scheduler to adjust the learning rate at specified milestones (γ). The model was trained for 100 epochs with a batch size of 6. For the 3D CNN, a kernel size of 5, stride of 1, and padding of 2 were used, while three slices from each plane (coronal, sagittal, axial) were stacked to increase channel depth. The transformer module employed 8 attention heads per layer, and binary cross-entropy loss was applied for binary classification tasks such as Alzheimer's Disease vs. normal controls. The performance of the classification method was evaluated using two key metrics: Area Under the Curve (AUC) and Accuracy (ACC). The implementation leveraged TensorFlow, along with the MedMNIST library for dataset management. Matplotlib and NumPy were used for visualization and preprocessing. Training on NVIDIA A100 GPUs took approximately 25 hours across all datasets.

C. Comparison study results

We conducted a comprehensive comparison of 3DViTMedNet against various benchmark models using the Med3D dataset. Specifically, we evaluated its performance against different configurations of 3D ResNet models, employing diverse convolutional approaches and model sizes (18 and 50). Additionally, we leveraged auto-sklearn and Auto-Keras, which employ automated machine learning techniques to optimize model selection and parameter tuning. Furthermore, to assess the comparative performance of our model against other transformer-based architectures, implementations of 3D ViT and 3D SWIN were evaluated.

The classification metrics presented in Table I offer valuable insights into the performance of 3DViTMedNet across datasets of varying modalities, class distributions, and characteristics. Notably, 3DViTMedNet achieves state-of-the-art performance for all classification metrics on the AdrenalMNIST3D dataset, and also demonstrates an improvement in accuracy on the SynapseMNIST3D dataset. While the performance on other datasets does not surpass state-of-the-art results, it remains competitive in most cases. These findings underscore the potential of hybrid transformer models in advancing 3D medical image classification tasks. The 3D ViT model consistently underperforms across all classification metrics for the evaluated datasets, which may be attributed to the limited availability of training data. In comparison, 3DViTMedNet significantly outperforms the pure ViT model in all cases and generally surpasses the 3D SWIN model as well. This demonstrates the strength of the hybrid transformer architecture in scenarios requiring both local and global feature extraction with limited data.

Dataset	Data Modality	Tasks (# Classes/Labels)	# Samples	Training /Validation / Test
OrganMNIST3D	Abdominal CT	Multi-Class (11)	1,742	971 / 161 / 610
NoduleMNIST3D	Chest CT	Binary-Class (2)	1,633	1,158 / 165 / 310
AdrenalMNIST3D	Shape from Abdominal CT	Binary-Class (2)	1,584	1,188 / 98 / 298
FractureMNIST3D	Chest CT	Multi-Class (3)	1,370	1,027 / 103 / 240
VesselMNIST3D	Shape from Brain MRA	Binary-Class (2)	1,908	1,335 / 191 / 382
SynapseMNIST3D	Electron Microscope	Binary-Class (2)	1,759	1,230 / 177 / 352

Fig. 3. The MedMNIST3D dataset [27] comprises six biomedical datasets of 3D images, each tailored for specific medical imaging tasks. The dataset includes various notations to denote task types, including MC (Multi-Class) and BC (Binary-Class).

TABLE I
3D MEDMNIST DATABASE RESULTS

3D Medical Image Classification Results												
Methods	OrganMNIST3D		NoduleMNIST3D		FractureMNIST3D		AdrenalMNIST3D		VesselMNIST3D		SynapseMNIST3D	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18 + 2.5D	0.977	0.788	0.838	0.835	0.587	0.451	0.718	0.772	0.748	0.846	0.634	0.696
ResNet-18 + 3D	**0.996**	**0.907**	0.863	0.844	0.712	0.508	0.827	0.721	0.874	0.877	0.820	0.745
ResNet-18 + ACS	0.994	0.900	0.873	0.847	0.714	0.497	0.839	0.754	**0.930**	**0.928**	0.705	0.722
ResNet-50 + 2.5D	0.974	0.769	0.835	0.848	0.552	0.397	0.732	0.763	0.751	0.877	0.669	0.735
ResNet-50 + 3D	0.994	0.883	0.875	0.847	0.725	0.494	0.828	0.745	0.907	0.918	**0.851**	0.795
ResNet-50 + ACS	0.994	0.889	0.886	0.841	**0.750**	**0.517**	0.828	0.758	0.912	0.858	0.719	0.709
auto-sklearn	0.977	0.814	**0.914**	**0.874**	0.628	0.453	0.828	0.802	0.910	0.915	0.631	0.730
AutoKeras	0.979	0.804	0.844	0.834	0.642	0.458	0.804	0.705	0.773	0.894	0.538	0.724
3DViT	0.636	0.325	0.795	0.790	0.572	0.412	0.549	0.769	0.616	0.856	0.600	0.731
3D SWIN	0.956	0.673	0.809	0.832	0.654	0.500	0.708	0.769	0.705	0.887	0.695	0.730
3DViTMedNet (Ours)	0.963	0.726	0.774	0.794	0.631	0.417	**0.883**	**0.846**	0.751	0.887	0.627	**0.798**

D. Interpretability of network focus

We visualized attention maps for 3DViTMedNet in figure 4 and ResNet models (using Grad-CAM for ResNet) in figure 5 respectively, to improve model interpretability. The figures provided represent the same input image (which can be visualised in 4). The ResNet model demonstrates concentrated attention in the center of each plane, focusing well on the adrenal gland. Meanwhile, 3DViTMedNet displays broader attention, effectively capturing global relationships while also increasing localized attention in the coronal plane, validating our hybrid approach to capturing both global and local features.

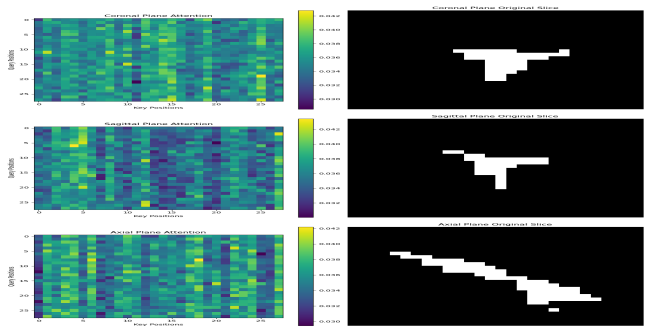

Fig. 4. AdrenalMNIST3D: 3DViTMedNet Attention Map in Comparison with Original Image.

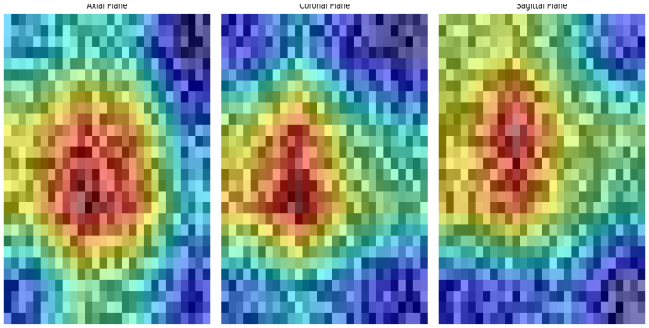

Fig. 5. AdrenalMNIST3D ResNet 50 + 3D Activation Map.

V. Conclusion and Future Work

This paper introduces 3DViTMedNet, a novel hybrid vision transformer model that incorporates both multi-slice and multi-plane extractions while processing 3D images directly. 3DViTMedNet is meticulously designed using 3D convolutional neural network (CNN) operations to extract representational features from the 3D datasets. These extracted features are subsequently passed through 2D CNNs, allowing the model to leverage pretrained networks before feeding the processed data into a vision transformer to capture global relationships. Through extensive experimentation, we evaluated the model against various benchmark models on the 3D MedMNIST database and other transformer-based architectures. Notably, 3DViTMedNet outperformed competing architectures on 2 out of the 6 datasets and achieved competitive results on the remaining datasets. These findings underscore the model's ability to excel in specific medical imaging tasks and demonstrate its potential to deliver strong performance in complex image classification challenges, particularly when optimized for certain dataset characteristics. A potential limitation of this work is the nature of the 3D MedMNIST database, which may pose challenges due to its low resolution of (28×28×28), which limits both human and model interpretability by providing insufficient detail for accurate diagnosis. Medical professionals typically rely on high-resolution images to discern fine anatomical structures, and deep learning models similarly require more spatial detail to learn meaningful patterns. Consequently, models trained on this dataset may struggle to extract features crucial for classification. Furthermore, transformer models, which excel at capturing global relationships, are likely hindered by the simplicity of this dataset, as it lacks the complex, non-local dependencies necessary for transformers to fully utilize their strengths. In contrast, CNNs, which focus on local feature extraction, are better suited to handle the limited spatial complexity, which may explain their superior performance compared to transformers in this case. For future research, we aim to develop a comprehensive, end-to-end imaging tool that integrates both segmentation and classification capabilities. This system would first perform segmentation to isolate relevant anatomical structures, followed by generating a classification result to facilitate more accurate and efficient medical diagnoses. This approach would streamline the workflow, enhancing the model's applicability across various medical imaging tasks by combining two critical processes into a unified framework.

References

[1] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 212–220.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014. [Online]. Available: https://arxiv.org/abs/1409.4842

[3] V. Wegmayr, S. Aitharaju, and J. Buhmann, "Classification of brain MRI with big data and deep 3D convolutional neural networks," in *Medical Imaging 2018: Computer-Aided Diagnosis*, N. Petrick and K. Mori, Eds., vol. 10575, International Society for Optics and Photonics. SPIE, 2018, p. 105751S. [Online]. Available: https://doi.org/10.1117/12.2293719

[4] J. Islam and Y. Zhang, "Brain mri analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informatics*, vol. 5, 05 2018.

[5] Kruthika, Rajeswari, and M. H.D., "Cbir system using capsule networks and 3d cnn for alzheimer's disease diagnosis," *Informatics in Medicine Unlocked*, vol. 14, 12 2018.

[6] X. W. Gao, R. Hui, and Z. Tian, "Classification of ct brain images based on deep learning networks," *Computer Methods and Programs in Biomedicine*, vol. 138, pp. 49–56, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260716305296

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[9] S. Korolev, A. Safiullin, M. Belyaev, and Y. Dodonova, "Residual and plain convolutional neural networks for 3d brain mri classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017, pp. 835–838.

[10] C. Yang, A. Rangarajan, and S. Ranka, "Visual explanations from deep 3d convolutional neural networks for alzheimer's disease classification," 2018.

[11] X. Gao, Y. Qian, and A. Gao, "Covid-vit: Classification of covid-19 from ct chest images based on vision transformer models," 2021.

[12] L. Zhang and Y. Wen, "A transformer-based framework for automatic covid19 diagnosis in chest cts," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 513–518.

[13] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, *Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities*.

Springer International Publishing, 2016, p. 478–486. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46723-8$_5$5

[14] E. Jun, S. Jeong, D.-W. Heo, and H.-I. Suk, "Medical transformer: Universal brain encoder for 3d mri analysis," 2021.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[16] 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," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers distillation through attention," 2021. [Online]. Available: https://arxiv.org/abs/2012.12877

[19] B. Gheflati and H. Rivaz, "Vision transformers for classification of breast ultrasound images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine 'I&' Biology Society (EMBC)*, 2022, pp. 480–483.

[20] D. Shome, T. Kar, S. N. Mohanty, P. Tiwari, K. Muhammad, A. AlTameem, Y. Zhang, and A. K. J. Saudagar, "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/21/11086

[21] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, Sep. 2022. [Online]. Available: https://doi.org/10.1145/3505244

[22] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523000634

[23] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," *Medical Image Analysis*, vol. 91, p. 103000, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523002608

[24] Y. Dai, Y. Gao, and F. Liu, "Transmed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, 2021. [Online]. Available: https://www.mdpi.com/2075-4418/11/8/1384

[25] C.-C. Hsu, G.-L. Chen, and M.-H. Wu, "Visual transformer with statistical test for covid-19 classification," 2021.

[26] J. Jang and D. Hwang, "M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 686–20 697.

[27] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, Jan. 2023. [Online]. Available: http://dx.doi.org/10.1038/s41597-022-01721-8

[28] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, "The liver tumor segmentation benchmark (lits)," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2019, pp. 146–164.

[29] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[30] D. Jin, P. Cao, A. Fedorov, S. Fulbari, Z. Gao, A. P. Harrison, M. Michalski, S. Napel, M. Pomerleau, R. M. Summers *et al.*, "Ribfrac: A challenge for automated rib fracture detection and classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2020, pp. 512–521.

[31] C. Qin, X. Huang, C. Liang, J. Ye, Z. Gao, X. Zhou, S. Tian, Y. Xie, B. Zhang, X. Ye *et al.*, "Intra: 3d intracranial aneurysm dataset for deep learning," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2930–2940, 2020.

[32] D. Wei, H. Wang, Z. H. Levine, W. Xie, T. Kroeger, D. Bock, J. Hegde, Y. Park, and H. S. Seung, "Mitoem dataset: large-scale 3d mitochondria instance segmentation from em images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.