3DViTMedNet: A Hybrid Transformer Architecture for 3D Medical Image Classification

Submitted as Research Report in SIT723/SIT746

SUBMISSION DATE

T2-2024

Spencer Gerontzos

STUDENT ID s224534582

COURSE - Bachelor of IT Honours (S470)

Supervised by: Dr. Son Tran

Acknowledgements

I would like to express my gratitude to Dr. Son Tran, for his invaluable guidance, continuous support, and encouragement throughout the course of this research project.
I also wish to extend my thanks to Justin Li from the School of Information Technology at Deakin University for providing the necessary resources and facilities that made this research possible.
Lastly, I am profoundly grateful to my family and friends for their support during this period.

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.

Contents

List of Figures

List of Tables

1 Introduction

In the realm of computer vision tasks, Convolutional Neural Networks (CNNs) have long been the cornerstone. This holds true especially in medical computer vision, where the predominant inputs are volumetric in nature, stemming from modalities such as X-ray, CT, MRI, and ultrasound.

The emergence of 3D inputs in the medical domain, coupled with the surge in computational resources, has spurred the development of deep learning models tailored to process such volumetric data. However, when considering approaches to handle these 3D inputs, researchers encounter a critical decision point.

One approach involves processing 3D inputs through 2D slices, leveraging pre-trained models trained on datasets like ImageNet to potentially enhance model performance. However, this method sacrifices spatial context crucial for interpreting medical images effectively.

Alternatively, directly processing 3D images while learning 3D representations presents its own challenges. The scarcity of publicly available medical data limits the complexity that can be learned, potentially hindering model performance.

Recently, the adaptation of transformers from the field of Natural Language Processing (NLP) to computer vision tasks has garnered attention. With their attention mechanism enabling a broader receptive field compared to CNNs, transformer-based models like Google's Vision Transformer (ViT) have demonstrated remarkable performance. However, they still fall short when compared to CNN methods, primarily due to the lack of extensive medical datasets.

In medical computer vision tasks, the importance of a large receptive field cannot be overstated. It allows for the identification of global patterns and relationships across an entire image, crucial for diagnosing diseases. Conversely, local context within an image, such as a specific anatomical structure or pathology, also plays a pivotal role. CNNs excel at capturing such local details owing to their high inductive bias.

In our study, we propose a novel approach: a hybrid architecture that combines both CNN and Transformer components 3DViTMedNet. By leveraging the strengths of both global and local receptive fields, this hybrid model presents an intriguing avenue for further exploration. 3DViTMedNet achieves competitive and State of the art (SOTA) results on the MedMNIST3D [63] database, offering a promising solution to the challenges posed by processing volumetric medical data.

1.1 Aim & Objectives

List of Aims

- Can 3DViTMedNet Justify the efficacy of hybrid transformers in 3D Medical image classification tasks
- Can 3DViTMedNet set new benchmarks for performance on the 3D Medical MNIST dataset?
- How does the integration of 3D and 2D feature representations in 3DViTMedNet impact the overall accuracy and interpretability of medical image classification?
- Can 3DViTMedNet effectively capture both local and global spatial features in volumetric medical data, and how does it compare to purely convolutional or transformer-based models in terms of computational efficiency and model complexity?

2 Background knowledge

This thesis will explore key concepts in deep learning architecture, focusing on foundational convolutional neural network (CNN) designs such as VGG and ResNet. Additionally, attention will be devoted to transformer architectures including Vision Transformer (ViT) and Swin Transformer (SWIN). By examining these architectures, this section aims to provide the reader a comprehensive understanding of the principles, strengths, and applications of architectures for 3D image classification tasks. We note that the aforementioned architectures above will be explained in 2D, however we will make relevant how the 2D architecture needs to be modified in order to accept volumetric inputs.

2.1 3D Convolution Neural Networks

Convolutional neural networks usually have 3 distinct layers, these being: (1) convolution layers, (2) pooling layers and (3) connected layers.
(1) convolutional layers:
3D implementation of a convolutional layer (1):

Convolutional layers in a CNN are responsible for extracting features from the input data. They achieve this by performing convolution operations between the input data and a kernel. The convolution operation involves sliding the filters over the input data and computing dot products between the filter weights and the values within a local receptive field of the input. These implementation is quite similar to that of a 2D CNN, where we simply increase the dimensionality of the convolutional layer from 2D to 3D .

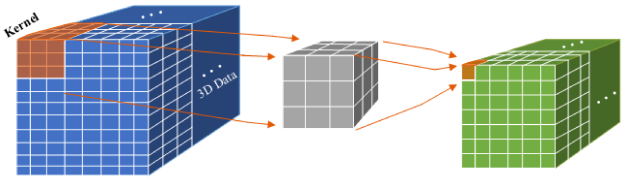

Figure 1: Visualization of a 3D Convolution Operation

Formulas:

The output of a convolutional layer is computed using the following formula:

$$Z^{\ell}_{i,j,k} = \sum_{a=0}^{n1-1} \sum_{b=0}^{n2-1} \sum_{c=0}^{n3-1} W^{\ell}_{a,b,c} \cdot X^{\ell-1}_{i+a,j+b,k+c} + b^{\ell}_{i,j,k}$$

where:

- $Z^{\ell}_{i,j,k}$ is the output activation at position (i, j, k) in layer ℓ.
- $W^{\ell}_{a,b,c}$ is the weight of the filter at position (a, b, c) in layer ℓ.
- $X^{\ell-1}_{i+a,j+b,k+c}$ is the input activation at position $(i+a, j+b, k+c)$ in the previous layer $\ell-1$.
- $b^{\ell}_{i,j,k}$ is the bias term associated with the output activation at position (i, j, k) in layer ℓ.

The output of the convolutional layer after applying a nonlinear activation function σ (such as ReLU) is given by:

$$A^{\ell}_{i,j,k} = \sigma(Z^{\ell}_{i,j,k})$$

where $A^{\ell}_{i,j,k}$ is the activated output at position (i, j, k) in layer ℓ.

3D Implementation of a Pooling Layer (2):

Pooling layers in a convolutional neural network (CNN) are used to reduce the spatial dimensions of the input data while retaining important features. They achieve this by applying a pooling operation to each local receptive field of the input data.

Formulas:

The most common type of pooling operation is max pooling, which selects the maximum value within each local receptive field. The output of a max pooling layer is computed using the following formula:

$$Y^{\ell}_{i,j,k} = \max_{a,b,c}\{X^{\ell-1}_{si+a,sj+b,sk+c}\}$$

where:

- $Y^{\ell}_{i,j,k}$ is the output activation at position (i, j, k) in layer ℓ.
- $X^{\ell-1}_{si+a,sj+b,sk+c}$ is the input activation at position $(si + a, sj + b, sk + c)$ in the previous layer $\ell - 1$, within the pooling region centered at (si, sj, sk).

The pooling operation can also be performed using average pooling, where the average value within each local receptive field is computed. The output of an average pooling layer is given by:

$$Y^{\ell}_{i,j,k} = \frac{1}{n_1 \times n_2 \times n_3} \sum_{a,b,c} X^{\ell-1}_{si+a,sj+b,sk+c}$$

where $n_1 \times n_2 \times n_3$ is the size of the pooling region.

Additional Techniques used to combat regularization

To ensure that the designed CNN does not fit the training data too closely (also known as overfitting) there are techniques that are used to combat this (which will be mentioned in later methods of classification)

Dropout

Dropout is a regularization technique commonly used in CNNs to prevent over fitting by randomly deactivating neurons during training. Let $x^{\ell}_{i,j,k}$ denote the output of the (i, j, k)th unit in layer ℓ. During training, dropout is applied to this unit with a certain probability p, settings its activation to zero when applied.

Batch Normalization

Batch normalization is a technique commonly used in 3D CNNs to stabilize and accelerate the training process by normalizing the activation's of each layer. Let $x_{i,j,k}^{\ell}$ denote the output of the (i, j, k)th unit in layer ℓ.

The batch normalization operation for this unit is computed as follows:

$$\hat{x}_{i,j,k}^{\ell} = \frac{x_{i,j,k}^{\ell} - \mu_{\ell}}{\sqrt{\sigma_{\ell}^2 + \epsilon}} \tag{1}$$

where:

- $\hat{x}_{i,j,k}^{\ell}$ is the normalized output of the (i, j, k)th unit in layer ℓ,
- $x_{i,j,k}^{\ell}$ is the original output of the (i, j, k)th unit in layer ℓ,
- μ_{ℓ} is the batch mean of layer ℓ,
- σ_{ℓ}^2 is the batch variance of layer ℓ,
- ϵ is a small constant (typically 10^{-5}) added for numerical stability.

After normalization, the normalized activation's are scaled by learnable parameters γ and β, and then passed through a non-linear activation function σ:

$$y_{i,j,k}^{\ell} = \sigma(\gamma_{\ell} \hat{x}_{i,j,k}^{\ell} + \beta_{\ell}) \tag{2}$$

where:

- γ_{ℓ} is the scaling parameter of layer ℓ,
- β_{ℓ} is the shift parameter of layer ℓ,
- σ is the activation function (e.g., ReLU).

2.1.1 VGGNet

VGG-16 is a convolutional neural network (CNN) specifically crafted for image classification duties [51]. Renowned for its straightforwardness and consistent architecture, VGG-16 is appreciated for its simplicity in comprehension and implementation. Typically comprising 16

layers, VGG-16 architecture encompasses 13 convolutional layers and 3 fully connected layers. These layers are structured into blocks, each housing several convolutional layers trailed by a max-pooling layer for down sampling purposes.

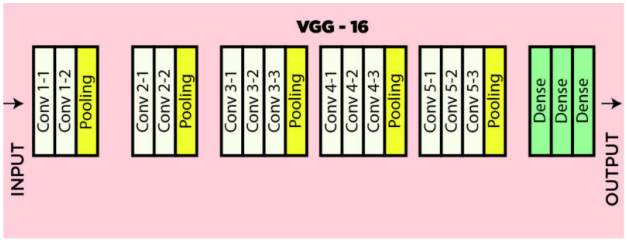

Figure 2: Visualization of VGGNet architecture

The VGG-16 architecture consists of several layers designed for image classification tasks:

- **Input Layer**: Accepts input images with dimensions (224, 224, 3).
- **Convolutional Layers**: Features two sets of consecutive convolutional layers, each with 64 filters of size 3x3, utilizing same padding to maintain spatial dimensions. This is followed by similar configurations with 128 filters and 256 filters.
- **Max Pooling Layers**: Includes max-pooling layers with a pool size of 2x2 and a stride of 2 after every two convolutional layers, aiding in downsampling.
- **Additional Convolutional Layers**: Integrates a stack of convolutional layers with filter size 3x3 after the previous stack, enhancing feature extraction.
- **Flattening**: Flattens the output feature map (7x7x512) into a vector of size 25,088 for further processing.
- **Fully Connected Layers**: Comprises three fully connected layers with ReLU activation. The first layer accepts the flattened input and outputs 4096 units, followed by two more layers with the same configuration. The final layer outputs probabilities for 1000 classes using softmax activation, corresponding to the classes in the ILSVRC challenge.
- **Remarks**: This architecture's modular design allows it to be adaptable to images and classes of varying sizes. With appropriate adjustments, such as resizing input images or modifying the output layer to accommodate different class numbers, VGG-16 can be effectively utilized across a wide range of image classification tasks.

VGG remains a powerful choice for image classification tasks, but its high accuracy comes with computational challenges. The original VGG-16 model, for instance, required around 2.5 weeks to train on an NVIDIA TITAN GPU for the ILSVRC challenge in 2014. Additionally, the model's weight size reached approximately 520MB, indicating inefficiency. The sheer number of parameters, around 140 million, also led to problems like exploding gradients. To address this issue, attention shifted to the ResNet architecture.

2.1.2 ResNet

As we increase the number of layers in a neural network, a common issue that arises is the vanishing/exploding gradient problem. This problem occurs when the gradients, which are propagated backward through the layers during training, become very small or very large. As a result, it becomes difficult for the network to update the weights effectively, leading to slow convergence. To mitigate these effects, activation functions like ReLU and techniques such as batch normalization are often used. However, He et al. developed a new architecture to directly address these issues [26].

The authors introduced the concept of skip connections, where activation's from one layer are connected to further layers by bypassing some intermediate layers. This approach helps combat the vanishing gradient problem by ensuring that the initial input to a network layer is combined with the output of that layer through element wise summation. By incorporating skip connections, the model maintains a more stable gradient flow during back propagation, which helps in alleviating the issue of decreasing gradients and enhances overall network performance. Below is a visualization of a skip connection, where we can see the input provided to the network is factored into the activation function after it has passed through network layers.

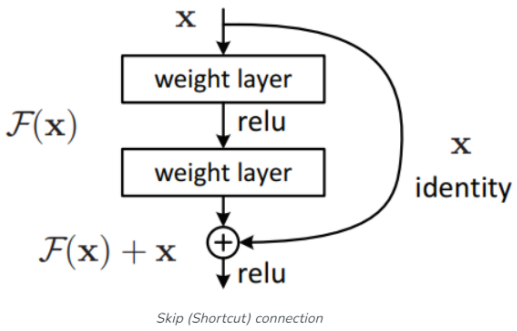

Figure 3: Visualization of a skip connection in practice

Using the idea of the VGGNet architecture mentioned earlier, He et al. build on this network

incorporating their skip connections to a 152 layer network, 8 times larger than VGG19, whilst having less parameters on the ImageNet Dataset. Below is a visualization of a ResNet-34 architecture which has 34 layers

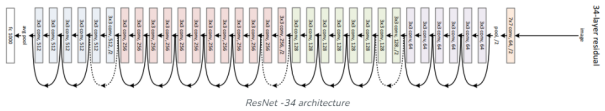

Figure 4: Visualization of the ResNet-34 architecture

2.2 Transformer Architectures for Image classification

In 2017, the Google Brain team introduced the transformer architecture, featuring the now renowned multi-head self-attention (MSA) mechanism [58]. This architecture has become foundational in the field of natural language processing (NLP), powering well-known models like GPT-2 and BERT. Although transformer architectures became standard for NLP tasks, their application to other domains was limited. However, in 2021, the Google Brain team adapted the transformer architecture for 2D image classification, achieving state-of-the-art results on the ImageNet dataset. Following this breakthrough, both Microsoft and Facebook developed their own transformer architectures for image classification, which we will explore in this section.

2.2.1 Transformers

The transformer architecture follows an encoder-decoder framework. The encoder processes the input and generates a sequence of representations, known as hidden layers. The decoder then takes the encoder's output as its input and produces the final output. Originally designed for NLP tasks, this architecture typically handles a 1D sequence of tokens as input and generates a single string as output. Unlike previous models, the transformer architecture can process information concurrently rather than sequentially, significantly improving efficiency. Below is a visualization of the transformer architecture.

Given that image classification tasks do not necessitate the generation of sequential outputs, the

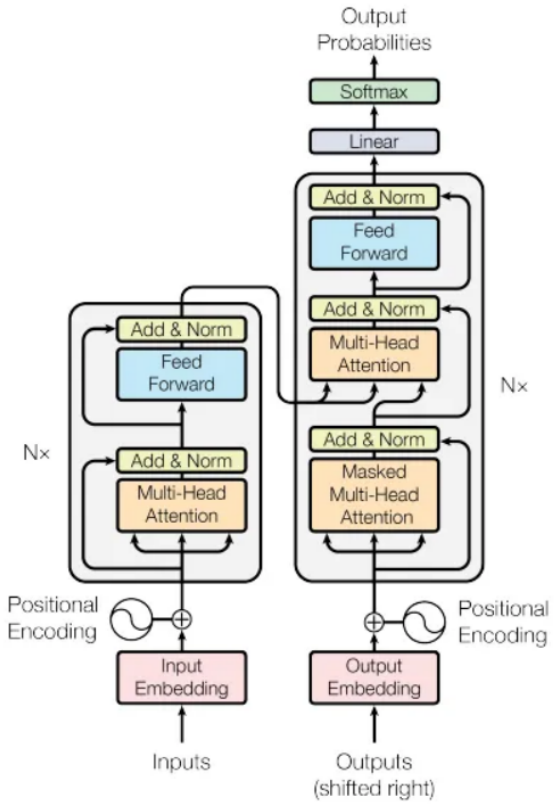

Figure 5: Visualization of the transformer architecture

decoder component of the architecture, which is specifically designed for this purpose, becomes irrelevant. Therefore, we will focus our attention on providing a detailed explanation of the encoder, which plays a central role in feature extraction and representation learning for image classification tasks.

Explanation of the Transformers Encoder Layer:

- **Input Embedding**:

  Input embedding's accept strings as input and use embedding layers to convert these strings into vectors that can be processed by the model. The process begins with tokenizing the input into words or subwords. Once tokenization is complete, these tokens are mapped to vectors through an embedding layer, such as Word2Vec.

- **Positional Encoding**:

  In the context of transformer architecture, the parallelization of processing a 1D sequence of tokens can potentially lead to the loss of relative positional understanding within the sequence. This understanding of relative position is crucial as adjacent tokens, words, or pixels provide contextual information to the model's input. To address this challenge, we augment each token with a positional embedding. This ensures that the model maintains a sense of relative positioning while processing the input sequence, enabling it to capture and leverage the contextual relationships between tokens effectively.

- **Stacking Encoders**:

  As illustrated in the visualization, both the encoders and decoders in the proposed architecture are structured with an 'Nx' multiplier adjacent to them. This signifies the number of layers utilized, with the authors opting for 6 layers in the paper. Each layer encompasses the Multi-Head Self-Attention (MSA) mechanism and a fully connected layer. Additionally, the architecture incorporates skip connections and layer normalization techniques to enhance regularization and mitigate the impact of the vanishing gradient effect.

- **Multi Head Self Attention**:

  Given an input sequence X, the multi-head self-attention mechanism operates as follows:

  1. **Linear Transformation:** We apply linear transformations to X to obtain the Query (Q), Key (K), and Value (V) vectors:

     $$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

     Here, W_Q, W_K, and W_V are learnable weight matrices.

  2. **Scaled Dot-Product Attention:** We compute the attention scores between Query (Q) and Key (K) vectors:

     $$\text{Attention}(Q, K) = \frac{QK^T}{\sqrt{d_k}}$$

     where d_k is the dimension of the Key vector.

  3. **Softmax and Weighted Sum:** We apply a softmax function to the attention scores to obtain attention weights:

     $$\text{Attention\_Weights} = \text{softmax}(\text{Attention}(Q, K))$$

     We then use these attention weights to compute a weighted sum of the Value (V) vectors:

     $$\text{Output} = \text{Attention\_Weights}V$$

  4. **Multi-Head Attention:** To enhance the model's ability to focus on different parts of the input sequence, we use multiple heads of attention. Each head computes its own Query, Key, and Value vectors using different learned weight matrices. The outputs of all heads are concatenated and linearly transformed to obtain the final output of the multi-head self-attention layer.

- **Feed Forward Neural Network**:

  The FFN consists of two linear (fully connected) layers responsible for transforming the input data. The first layer expands the input dimension to a larger space, while the second

layer projects it back to the original input dimension. To introduce non-linearity and enhance the model's expressive power, a Rectified Linear Unit (ReLU) activation function is applied between these two linear layers.

2.2.2 ViT

In 2021, the Google Brain team introduced a groundbreaking image classification architecture called Vision Transformer (ViT), which adapts the transformer architecture with slight modifications for handling images [19]. Given the computational complexity of computing attention for each pixel in an image (a quadratic task), ViT employs a clever strategy: it divides the image into smaller patches, typically 16x16 pixels, and applies a linear projection followed by positional embedding to each patch. This approach maintains the essence of the original transformer architecture while accommodating the unique characteristics of image inputs.

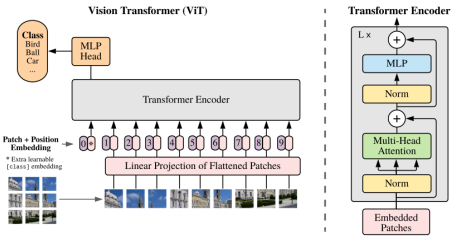

Figure 6: Visualization of the ViT architecture

Additionally, ViT incorporates a normalization layer before the Multi-Head Self-Attention (MSA) mechanism, enhancing its stability and performance. Unlike transformer architectures used in natural language processing tasks, ViT aims to produce classification results. To achieve this, the feature representations obtained from each patch are fed into a vanilla Multi-Layer Perceptron (MLP) block, which generates the final classification output.

2.2.3 SWIN

In response to the quadratic complexity challenge, Microsoft developers introduced groundbreaking adjustments to the transformer architecture, giving rise to the Swin architecture, leveraging shifted window attention and patch merging techniques [41]. Similar to ViT, Swin employs embedding to tokenize inputs, but with a notable departure: it operates on 4x4 patches instead of ViT's 16x16 patches.

The Swin transformer mechanism consists of two distinct transformer blocks. The first block features Window Multi-Head Self-Attention (W-MSA), which divides the input patch into quarters, allowing self-attention to be computed for each input pixel within these quarters. The rationale behind this approach lies in the assumption that pixels located far apart within an image are often unrelated. However, when crucial spatial and relational context is involved, such as a central object, dividing the input into quarters may risk losing such context. This motivates the second transformer block, the Shifted Window MSA (SW-MSA).

The SW-MSA addresses this challenge by shifting patches by half their original size, such that if the patch size is 4x4, it's shifted by 2x2 to the left and upwards. However, this introduces a new issue: some patches may extend beyond the defined window boundaries. To mitigate this, the authors propose a rolling mechanism, where patches exiting the window boundaries reappear in a different location within the window. For example, a patch shifted from the top-left quarter would reappear in the bottom-right quarter after the shift.

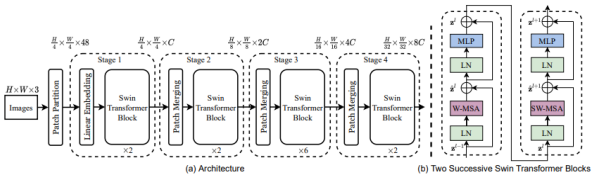

Figure 7: Visualization of the Swin Transformer architecture

This innovative rolling mechanism ensures that the transformer mechanism can be computed in linear runtime, a significant breakthrough introduced by Microsoft researchers. It enables efficient and effective processing of image inputs while preserving spatial and relational context.

Whilst the results are not state of the art like ViT, Swin transformers require far less data and resources to produce comparable results on ImageNet

2.3 Summary

This section has provided a concise introduction and refresher on key concepts in the field of image classification. We have reviewed fundamental ideas, focusing on convolutional neural networks (CNNs) and their components, as well as notable architectures like VGGNet and ResNet. Additionally, we explored transformer architectures, delving into models such as SWIN and ViT, and discussed the pivotal self-attention mechanism that underpins transformers. With

this foundational knowledge, we are well-equipped to examine proposed architectures for 3D image classification, leveraging our understanding of these essential concepts.

3 Literature Review

In recent years, the field of medical image classification has made significant strides, driven by advances in deep learning, particularly Convolutional Neural Networks (CNNs). However, despite their successes, CNNs face inherent limitations when it comes to modeling long-range dependencies and capturing global context in complex, high-dimensional medical data such as 3D volumes (e.g., Computed Tomography or Magnetic Resonance Imaging scans). As medical imaging increasingly demands more sophisticated models that can efficiently handle 3D data while retaining global context, the emergence of Vision Transformers (ViTs) and their 3D variants offers a promising alternative. Unlike traditional CNNs, transformers rely on self-attention mechanisms, which allow them to model intricate spatial relationships across an image or 3D volume more effectively.

The purpose of this literature review is to provide a comprehensive overview of existing research on 3D ViT architectures applied to medical image classification. By examining recent developments, this review aims to identify the strengths and limitations of these models and highlight areas for future research.

The importance of this topic is underscored by the increasing reliance on accurate and efficient medical image classification in critical areas such as disease diagnosis, treatment planning, and patient monitoring. As 3D medical imaging technologies such as MRI and CT scanning continue to evolve, so too must the computational methods used to interpret them. 3D ViT architectures hold the potential to transform how medical images are processed and understood, offering superior accuracy, better feature extraction capabilities, and improved handling of high-dimensional data compared to traditional architectures. Given the potentially life-saving implications of advances in medical image classification, it is crucial to explore the current landscape of 3D ViT, their performance, and their application to clinical tasks.

The literature for this review was selected through a comprehensive search of recent conference proceedings from 2019 to 2024, emphasizing cutting-edge advancements in 3D computer vision models for medical image classification. This timeframe was chosen to ensure the inclusion of the most current research, aligning with the rapid evolution of 3D medical image analysis techniques, particularly given the growing interest in hybrid architectures combining convolutional neural networks (CNNs) and transformers.

Keywords used during the search process included terms such as “3D CNN,” “3D Vision Transformers,” “hybrid models,” “medical image classification,” and “volumetric data analysis.” These terms were targeted to capture a wide range of methodologies relevant to both CNN-based and transformer-based approaches in medical imaging.

Key conferences such as ICCV [3], CVPR [2], MICCAI [4], and NeurIPS [1] were prioritized due to their influence and high impact in the fields of computer vision and machine learning. These conferences consistently publish cutting-edge research, providing a rich source of the latest advancements in 3D vision models and their applications to medical data. By focusing on these conferences, the review ensures that the selected studies represent the most relevant and impactful contributions to the field.

The scope and organizational structure of this literature review will focus on key advancements in 3D computer vision models for medical image classification. The review will be divided into two main sections: one exploring architectures based on 3D CNNs, and the other discussing 3D Transformer models. Additionally, hybrid approaches that integrate the strengths of both model families will be examined, providing a comprehensive overview of the current state of the field.

3.1 CNN in Medical Image Classification

The research of Nie et al. offers valuable insights into predicting survival outcomes for patients with high-grade gliomas by leveraging deep learning frameworks to extract features from multi-modal brain imaging data, including T1 MRI, functional MRI (fMRI), and diffusion tensor imaging (DTI) [45]. The authors employ 3D CNNs and propose an innovative architecture capable of processing multi-channel data to effectively learn supervised features from the diverse imaging modalities. These learned features, when integrated with key clinical information, are utilized to train a support vector machine (SVM) to predict patient overall survival (OS) times, classifying them into long or short survival groups.

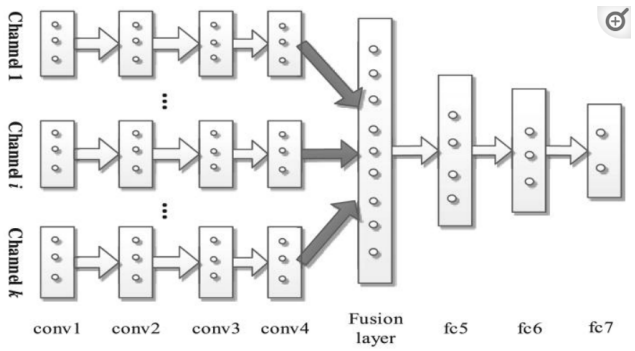

Figure 8: Architecture of the CNN for feature extraction from multi-channel data

This study underscores the efficacy of CNN models as powerful feature encoders and extractors.

Wegmayr et al. combined both Alzheimer Disease Neuroimaging Initiative (ADNI) [5] and Australian Imaging Biomarkers and Lifestyle Study of Ageing (AIBL) [9] datasets to produce one of the largest databases of both subjects and images [59]. Where the authors' work differs from above previously mentioned architectures is the implementation of the first convolutional layer 9. Authors use 3 different kernel sizes, with the design philosophy that it will capture various features across varying scales. This idea enables a CNN to understand an input comprehensively.

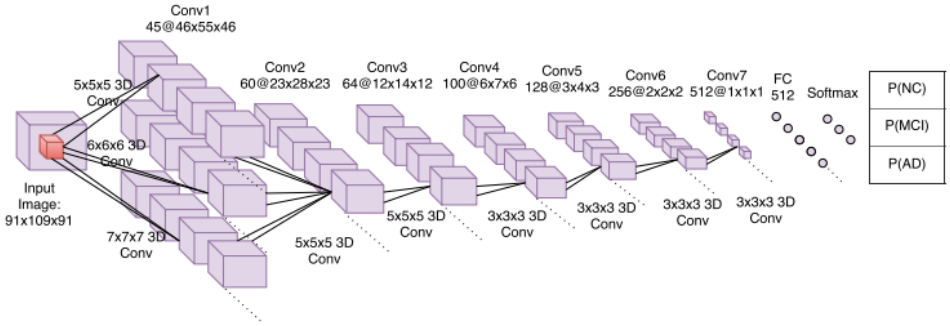

Figure 9: Visualization of the multi scale convolutional layer deployed by Wegmayr et al. [59]

Wegmayr et al. [59] utilized an ensemble technique by averaging the class probabilities from the last four epochs of training. This methodology aimed to enhance model robustness by incorporating insights gathered across multiple epochs. The authors chose to solely assess the accuracy of their model against other established architectures in the field, despite having one of the largest datasets comprising over 20,000 subjects. It was anticipated that their model, benefiting from such extensive data, would outperform the benchmark models. However, Wegmayr et al. [59] pointed out the potential risks of over-fitting in their comparison, and they did not elaborate on the normalization methods or the approach to handling duplicate subjects.

Also using an ensemble architecture, Islam and Zhang where able to produce an ensemble system of Deep CNNs to similarly, perform Alzheimer Disease (AD) classification on the OASIS dataset [32]. Where [38] utilised 3D convolutions, Islam and Zhang used an ensemble of 3 Models to respectively classify the Coronal, Saggital and Axial planes of the 3D dataset, where they perform a max voting ensemble, combining the results of the 3 models to produce the final classification of the dataset 10.

The authors note the effect that the density of this network has regarding regularization, where due to the shear number of connections and layers, it prevents overfitting. The authors also propose slightly adjusted models: M4 and M5 that are also used as a comparison towards the main models: M1, M2 and M3. The authors first evaluate each of the respective models based on Recall, Precision and F1-score. They also show the support of each of these classes which

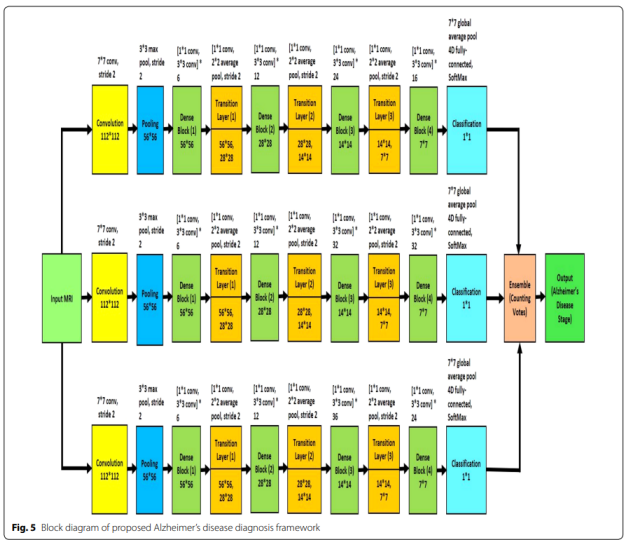

Figure 10: Visualization of the ensemble architecture proposed by Islam and Zhang [38]

helps us understand some of the results. Due to the scarcity of data in this field, we note that the test dataset only has a total of 88 images to classify against, where over 80% of the images are apart of the NC class of images, where the remaining 15 images are not uniformly distributed about the remaining classes. This is clear when we review the evaluation of each of the individual models, where nearly all models generate precision's larger than 85% and recalls larger than 90% (excluding model 4). The authors also produce a comparison with other benchmark models which achieves notably higher results than the counterparts of Resnet, ADNet and Inception -V4. These metrics presented are the mean of the results mentioned above, where as we have mentioned, the NC class would heavily skew these results. Instead of generating the average of these metrics and comparing them to benchmark models, it would have been great to see the evaluation metrics these benchmark models also proposed.

Korlev et al. provide valuable insights into the performance of benchmark CNN models for 3D AD classification [38]. In their study, both VGGNet and ResNet architectures were trained on the ADNI dataset. While their models achieved competitive results on binary classification tasks, a potential limitation of the architecture was revealed through the heat map analysis presented 11.

The network's attention was predominantly focused on the hippocampus, a highly localized region, which is only one of several critical areas used by medical professionals to diagnose AD. This observation highlights a key limitation of CNNs: their tendency to focus on local features, which can hinder the model's ability to capture broader, global relationships across the entire image. This limitation underscores the need for architectural advancements to better represent complex medical imaging data.

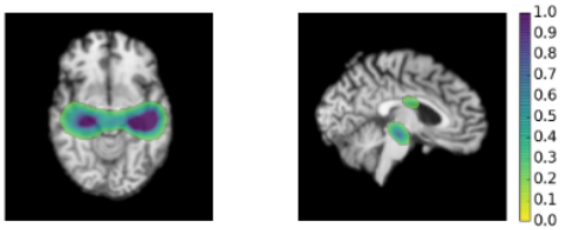

Fig. 3. Network attention areas for Normal Cohort MRI. (Axial and sagittal view)

Figure 11: Visualization of the network attention in convnet models proposed by Korlev et al. [38]

Kruthika et al. propose CBIR system, utilizing a 3D Capsule Network (CapsNet), 3D CNN, and a pre-trained 3D auto encoder for early AD detection [39]. The study found that an ensemble method combining 3D CapsNet with 3D CNN and a 3D auto encoder outperformed traditional Deep CNN models, achieving an AD classification accuracy of up to 98.42% on the ADNI dataset. Kruthika et al. [39] suggest that CapsNet is a promising technique for image classification and that future research with more robust computation resources and refined CapsNet architectures may yield even better results 12.

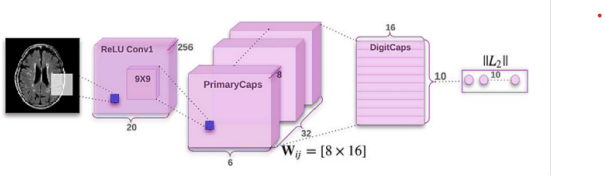

Figure 12: Visualization of the CBIR System utilizing Capsule Networks by Kruthika et al. [39]

Although the capsule network outperforms CNN models in 3D AD detection, it still struggles to capture spatial relationships effectively. This issue arises because 3D convolutional layers downsample the feature space, and flattening operations before the fully connected layer can lead to loss of spatial information.

Continuing with the topic of CNN models for AD detection, Gao et al. conducted a recent study that demonstrated the effectiveness of fusing 2D and 3D networks, achieving an impressive 87.7% accuracy in classifying AD, lesions, and normal aging using 285 volumetric CT head scans from the Navy General Hospital in China [23]. This fusion architecture integrates 2D images along spatial axial directions with 3D segmented blocks, combining their outputs through the average of Softmax scores from both networks 13.

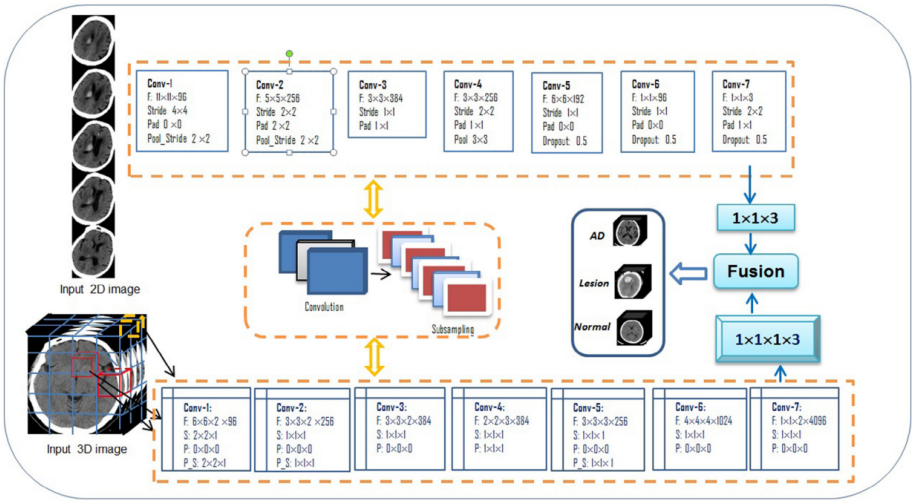

Fig. 2 – The fusion of both 2D and 3D CNNs for GT images.

Figure 13: Visualization of the fusion System of Gao et al. [23]

The proposed approach not only surpasses traditional 2D CNN models but also outperforms several state-of-the-art hand-crafted feature extraction methods by approximately 4%, highlighting the efficacy of this hybrid deep learning framework for medical image classification tasks.

Yang et al. [62]. employed 3D versions of ResNet [26] and VGGNet [51] to classify AD using MRI data from the ADNI dataset. While their classification results were competitive, the study's primary contributions centered around improving and interpreting the 3D CNN. The model was trained to classify AD by identifying key AD-related features, marking a significant advancement in understanding the behavior of each layer within a 3D CNN.

The authors introduced three techniques for visual inspection to enhance interpretability: sensitivity analysis, 3D class activation mapping (CAM), and 3D gradient-weighted class activation mapping (Grad-CAM) respectively [48]. The results showed that sensitivity analysis was effective at identifying local relationships, though it struggled to detect segments of the cerebral cortex—an expected limitation due to the local nature of CNN architectures. Both Grad-CAM and CAM, on the other hand, were able to loosely identify relationships with the cerebral cortex. These inspection methods inform architectural decisions for the 3D CNN. This study displays that CNN's struggle with identifying global relationships in complex imaging tasks.

Feng et al. significantly improved performance on binary classification tasks, achieving nearly 95% accuracy on the ADNI dataset through their innovative use of a fully stacked bi-directional long short-term memory (FSBi-LSTM) integrated with a 3D CNN architecture [21]. Traditional fully connected layers in a 3D CNN can lose spatial context due to their 1D input constraint. However, by incorporating LSTM layers [52] before the fully connected layers, the authors were able to retain both high-level semantic and spatial information, resulting in enhanced

performance. While they achieved notable accuracy for both binary classification and progressive mild cognitive impairment (pMCI), their model struggled with stable mild cognitive impairment (sMCI), achieving only about 65% accuracy.

In contrast to prior research on the ADNI dataset, Feng et al. [21] segmented the brain into 93 regions of interest and conducted classification on each of these regions independently. Their findings confidently highlighted the cerebral cortex as a critical component in AD classification.

AD modeling plays a crucial role in advancing early diagnosis and treatment strategies, particularly as it highlights the limitations of CNNs in complex imaging tasks. Despite their success in capturing local features, CNNs often struggle to identify global context across images, which is essential for comprehensive medical image analysis. This limitation emphasises the need for innovative architectural improvements in deep learning to better address the intricate spatial relationships present in brain imaging for AD detection.

3.2 ViTs in Medical Image Classification

3.2.1 Vanilla ViTs in Medical Image Classification

ViTs have gained attention in medical image classification due to their ability to model long-range dependencies and capture global context within images through self-attention mechanisms. Unlike traditional CNNs, which are limited by their localised receptive fields, ViTs intend to processing complex patterns across an entire image.

Since the release of the ViT architecture, many questions have emerged regarding its potential to replace traditional CNNs. One key question is whether there should be a complete shift away from CNNs. Matsoukas et al. conducted a comparative study between a ResNet-50 model and a DeiT-s model, both of which share similar parameters, memory and computational costs [43]. They trained these models using three different strategies: random initialization, transfer learning pre-trained on ImageNet [16], and self-supervised learning (SSL) on the target dataset using DINO [13]. The study focused on the APTOS 2019 [7], ISIC 2019 [31], and CBIS-DDSM [12] medical datasets respectivley.

Their findings revealed that when using random initialization, ViTs performed worse than CNNs. However, when transfer learning was applied, the results were similar for both models, with the ViT outperforming CNNs in two out of three datasets. Moreover, ViTs demonstrated superior performance when self-supervised learning on the target dataset was employed. The authors concluded that ViTs are indeed viable replacements for CNNs, particularly when utilizing transfer learning or self-supervised learning approaches.

Building on similar efforts, Geflati et al. evaluated pure ViT models with a linear classifier, comparing them against ViT models with an MLP head on breast ultrasound (US) images [24]. The authors compared these ViT models to state-of-the-art CNN architectures, with all models undergoing transfer learning to address the low inductive bias inherent in ViT architectures. The results were promising, as the ViT models matched and often outperformed the state-of-the-art ResNet models on the combined BUSI+B and BUSI [49] datasets. The authors suggest that in ultrasound imaging, there may be a stronger emphasis on spatial relationships, where large-scale dependencies between image patches play a more significant role, contributing to the success of ViTs in this particular modality.

The COVID-Transformer of Shome et al. utilizes a ViT architecture to differentiate between COVID-19 and non-COVID cases using chest X-ray (CXR) images [50]. To overcome the challenge of limited data, the authors created a balanced dataset of 30,000 CXR images for multi-class classification and 20,000 images for binary classification. The multi-class dataset comprises three categories: COVID-19 (for infected patients), normal (for healthy individuals), and pneumonia (for patients with viral or bacterial pneumonia), consistent with the classifications commonly seen in the literature.

The model was fine-tuned on this dataset, incorporating a custom Multi Layer Perceptron (MLP) block on top of the ViT for classification. To further enhance interpretability, the COVID-Transformer employed GradCAM maps to visualize the lung regions most significant for disease prediction and progression. These visualizations reveal that the model's attention effectively covers the entire input image, offering insights into how the ViT detects and classifies lung abnormalities linked to COVID-19. In terms of performance, the COVID-Transformer was benchmarked against popular CNN architectures, including EfficientNetB0 [54], InceptionV3 [53], ResNet50 [27], and MobileNetV3 [28], and outperformed all of them across all classification metrics.

Unlike previously mentioned architectures for 2D image classification, which rely on transfer learning techniques to improve the model performance. Gao et al. [22] showcased an innovative use of a pure ViT architecture during their participation in the MIA COVID-19 Challenge [47] . Their method involved transforming volumetric data into 2D slices, which were then directly input into the ViT for classification. To achieve classification, Gao et al. [22] employed a voting system for each slice. The ViT would increment the number of votes for slices classified as indicative of COVID-19. If a significant portion, say 75%, of slices were classified as COVID-19-positive, the entire volume would be similarly classified. One challenge faced by the researchers was the significant variation in the number of slices across chest CT scans in the dataset. To address this, they standardized the 3D volume to contain 32 slices. If a volume had more than 32 slices, they processed it in separate 'sub-volumes', combining the votes afterward. Despite its simplicity, this approach achieved an impressive accuracy of 76.6%, surpassing Densenet [30] by a notable 3% margin 14.

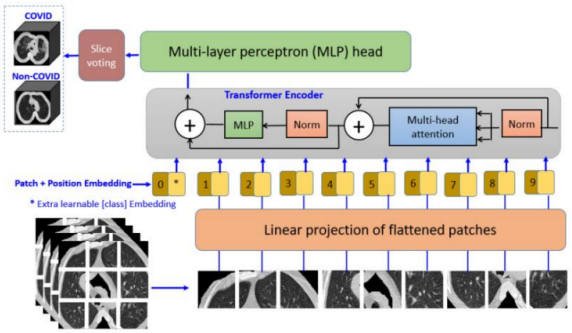

Figure 14: Visualization of Gao et al. proposed architecture [22]

A key challenge faced by the researchers was the variation in the number of slices across CT scans in the dataset. To address this, they standardized the 3D volumes to contain 32 slices. For volumes exceeding this limit, they divided the data into separate ”sub-volumes,” aggregating votes from all sub-volumes for the final classification. Despite the simplicity of this approach, it achieved an impressive accuracy of 76.6%, outperforming Dense net by a 3% margin. However, it is important to note that this method sacrifices some 3D representational features by processing the data in 2D slices. Additionally, due to the nature of ViTs, which process 1D sequences of tokens, spatial context can be somewhat compromised, as patch and position embedding's only retain limited spatial information. Moreover, the ViT's self-attention mechanism has quadratic complexity, which can impact computational efficiency when handling large-scale medical data. [19].

Instead of using ViT such as [22] when attempting to compete in the MIA COVID-19 Challenge, Zhang and Wen use the SWIN architecture [65] to perform the classifcation. The researchers first performed image segmentation on the chest CT in order to segment the lungs, this is then followed by the aforementioned SWIN transformer [41] architecture to classify the input. Researchers compared these results to architectures: BiT [37] and EfficientNetV2 [55], were they found the SWIN model produced the best F1 Score after evaluation at 93.5% 15.

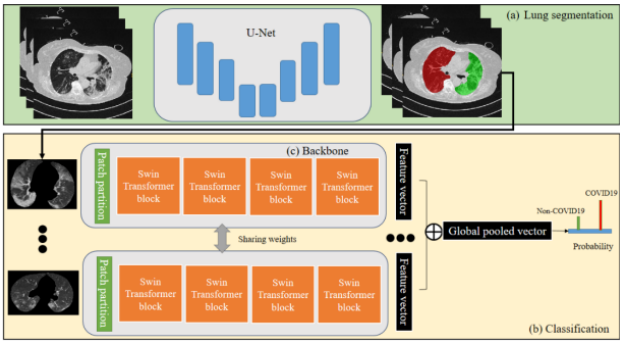

Figure 15: Visualization of Zhang and Wen's proposed architecture

However it was noted that the SWIN architecture produced 88 million parameters, whilst EfficientNetV2 produced 55 million parameters and a near identical F1-score. This highlights that there exists speed-accuracy trade offs when discussing model comparisons.

Vanilla ViT architectures in medical image classification face challenges due to their low inductive bias, which makes them less effective at capturing inherent spatial structures compared to convolutional models. This lack of inductive bias leads to difficulties in identifying local relationships within medical images, which are crucial for recognizing fine-grained features, such as lesions or abnormalities. As a result, pure ViTs may struggle to accurately model the complex patterns needed for tasks like medical image classification without the integration of additional mechanisms to enhance local feature extraction. Research has increasingly focused on hybrid ViT models that combine the strengths of transformers with convolutional layers to address the limitations of pure transformer architectures. These hybrid models aim to improve local feature extraction while leveraging the global attention capabilities of transformers, making them more suitable for medical image classification tasks. This approach will be explored in further detail below.

3.2.2 Hyrbrid ViTs in Medical Image Classification

Despite ViTs' strength in capturing global contextual representations, the self-attention mechanism tends to overlook important low-level details. To address this limitation, CNN-Transformer hybrid approaches have been developed, combining the local feature extraction capabilities of CNNs with the long-range dependency modeling of Transformers. This integration allows for encoding both local and global features more effectively, improving performance on tasks requiring detailed spatial representation.

Dai et al. introduced TransMed, an innovative multi-modal image classification approach utilizing a hybrid transformer architecture while maintaining the use of volumetric slices for analysis [15]. In their method, input modalities are fused, and a standard 2D ResNet is employed to downsample the images and extract local features. Unlike conventional architectures, the authors utilized a modified DeiT architecture [57], which processes the ResNet output to predict the most likely class. This modification eliminates both the linear projection layer—since the ResNet output is already vectorized—and the distillation token, a key component of DeiT's 'student-teacher' model. 16

Using the PGT dataset [6], which consists of various MRI modalities of the head and neck, the authors progressively adjusted their architecture to increase the number of parameters. TransMed achieved an impressive mean classification accuracy of 88.9%, with the highest accuracy across all classes. The confusion matrix highlighted that the BCA class had the lowest classification accuracy, with a 79% true positive rate. TransMed also demonstrated strong performance on

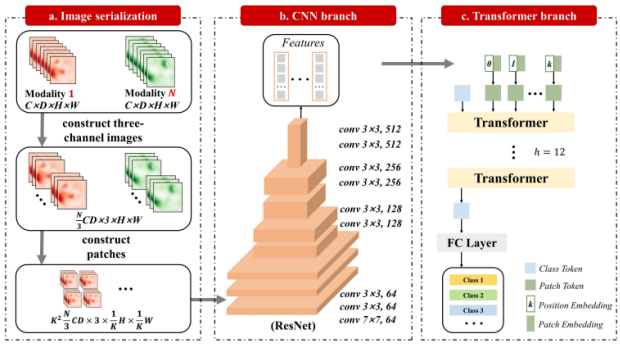

Figure 16: Visualization of Dai et al. proposed Transmed Architecture [15]

the MRNet dataset [10] for knee scans. Although this approach significantly outperforms other models using modality fusion techniques, its benefit may be limited for research that focuses on a single imaging modality. Nevertheless, the study provides valuable insights into the impact of a modified DeiT architecture in multi-modal medical image classification.

Instead of processing 3D medical images via the use of 2D slices, Tang et al. propose a SOTA self supervised pre-training of SWIN transformers for Medical imaging tasks [56]. Differently to the research above, where they utilize 'vanilla 2D' transformers, this article proposes SWIN-UNETR, a pre-trained SWIN transformer encoder. The main difference in this architecture is specifically the implementation of a 3D SWIN transformer, where the tokens and patches are in 3D. Furthermore, they apply a CNN decoder to create a 'U-Net' like architecture (commonly used in segmentation tasks) to produce leading results in the BTCV challenge on multi-organ segmentation, this architecture is an extension from previous work [25]. We note that a strength of this model was due the self supervised pre-training techniques (that are not relevant to this review of literature), where researchers were able to use unlabelled data to train the transformer encoder on useful features. **??**

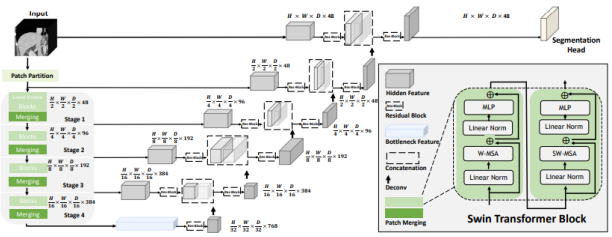

Figure 17: Visualization of Tang et al. proposed architecture [56]

Following on from using the idea of skip connections, Dai et al. have developed an innovative image classifier designed for the fMRI modality [14]. Unlike other approaches that may simply feed outputs from a CNN into a ViT architecture (or vice versa), the authors integrated a ResNet-like architecture where the residual connections utilize transformers. Given the 3D nature of the inputs, the authors addressed the challenge of the quadratic complexity of multi-head self-attention (MSA) mechanisms by incorporating Shallow Global Attention (SGA). In this

approach, instead of using MSA, they employed two MLP blocks for 'spatial mixing'. The spatial-mixing block allows communication between different spatial locations by operating independently on each channel, while the channel-mixing block facilitates communication between different channels. These two types of blocks are interleaved to enable interaction among input dimensions. This operation runs with linear complexity, providing a reasonable runtime for experiments. Once the input is sufficiently down sampled and reaches a deep layer, the authors then use MSA to effectively capture the global correlations between distant brain regions. 18

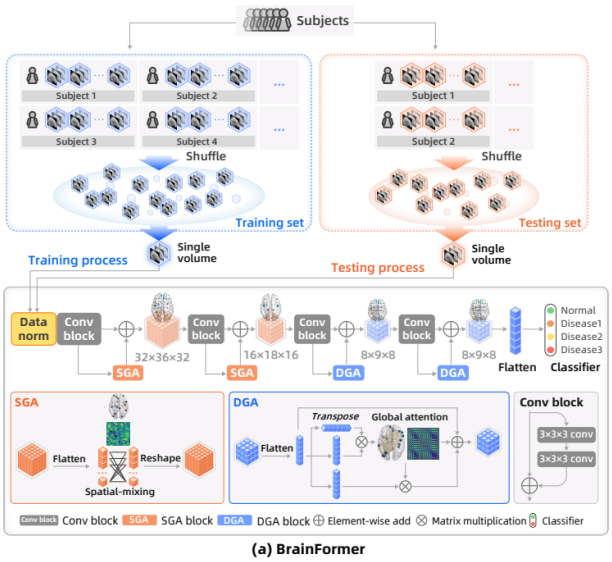

Figure 18: Visualization of Dai et al. ResNet architecture [14]

This hybrid architecture was tested on multiple datasets, including MPILMBB [40], ADNI [5], ABIDE [17], ADHD [44], and ECHO [20], consistently outperforming benchmark architectures on accuracy metrics. Additionally, the authors evaluated the Brainformer architecture [18] across various methods, including different ratios of shallow attention blocks to global MSA blocks, and comparisons of accuracy's without data normalization, with shallow attention, global attention, and without transformers altogether. This highlights the effectiveness of the ResNet-inspired architecture. However, a potential concern with this paper is the lack of efficiency metrics. Given the computational expense of transformers, it would be beneficial to see data on the number of parameters, hardware used for training, and the time taken for training and inference.

Hsu et al. proposed two innovative approaches for classifying COVID-19 cases, focusing on both 2D slices and volumetric data processing. The first approach utilizes the SWIN Transformer architecture [29]. Processing selected slices from the middle of the CT scans, where significant COVID-19 symptoms are most visible. After applying outlier removal, the researchers conducted a Wilcoxon signed-rank test [61] to assess statistical differences between the distributions of COVID-19 positive and negative slices. This test provided a measure of whether the observed differences were statistically significant, and the approach yielded impressive results, achieving an F1 score of nearly 92%. Notably, this method avoids the need for segmentation prior to

classification, showcasing a favorable speed-accuracy trade-off compared to other segmentation-dependent methods.

In the second approach, Hsu et al. [29] introduced a hybrid model. They used a ResNet-50 architecture, omitting global average pooling to preserve spatial context, to extract feature maps from the input slices. Unique to their method, SWIN Transformers were employed to process each individual feature slice, and an additional SWIN Transformer was applied to the feature maps of multiple slices to capture temporal or spatial dependencies between slices. This novel combination of Within-Slice and Between-Slice Transformers allowed the model to leverage contextual information across slices, significantly enhancing performance 19. The approach demonstrated outstanding accuracy, with approximately 93% accuracy and F1 scores, underscoring the potential of SWIN Transformers for improving medical image classification.

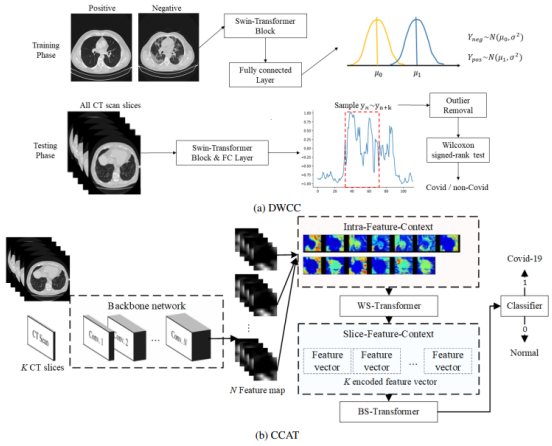

Figure 19: Visualization of both Hsu et al. proposed architectures [29]

Jun et al. utilize a multi plane multi slice approach [36] similar to the above literature. While the approach presented in the paper is promising, it covers a wide range of prediction tasks, including classification, regression, and segmentation 20. As a result, despite using several benchmark models—both pre-trained and non-pre-trained, via self-supervision or full supervision—the authors primarily reported only mAUC scores (achieving the best result of 0.8347 ± 0.0072). However, the paper lacks a broader set of evaluation metrics, such as F1 score, precision, or recall, which would provide a more comprehensive assessment.

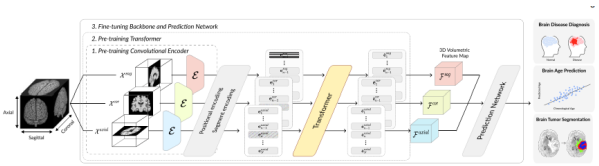

Figure 20: Visualization of Jun et al. proposed architecture [36]

One notable aspect of the study, which sets it apart from others, is its comparison of supervised learning performance relative to the amount of available training data. The authors demonstrate that using 100% of the available data in a fully supervised approach yields an mAUC of 0.7728 ±

0.0077. Interestingly, they show that by applying transfer learning, better results can be achieved using only 50% of the training data. Additionally, the study examines the impact of including a transformer in the architecture. Results indicate that incorporating the transformer led to nearly a 1% improvement in mAUC, underscoring the potential benefits of combining transformers with CNNs in this domain.

Building on the approach of using volumetric slices, Jang and Hwang proposed a hybrid transformer architecture for AD classification using the ADNI dataset, known as M3T [33]. Recognizing the localized brain structural changes associated with AD, they integrated CNNs to capture these local features, while employing transformers to model long-range dependencies using the self-attention mechanism.

Their method begins by applying a 3D CNN to the brain images, with two 3D convolutional layers utilizing square kernels of size 5 to extract 3D representational features. These 3D features were then sliced along the Coronal, Sagittal, and Axial planes. For each plane, a dedicated 2D convolutional block was applied to further refine the extracted features for input into the transformer model.

The features from each slice were subsequently fed into a transformer encoder, which processed the data and produced the final classification. This multi-plane, multi-slice architecture, referred to as M3T, demonstrated remarkable performance, achieving excellent AUC and accuracy metrics across multiple datasets, including ADNI [5], AIBL [9], and OASIS [42].

While this study provided a strong benchmark comparison against other models, the authors noted a need for additional evaluation metrics, such as F1 scores and precision, to further assess the model's effectiveness across various classification scenarios.

Across both pure and hybrid models, a common trend emerges: the processing of volumetric data through 2D slices, whether sampled from specific planes or a combination of planes. However, these approaches tend to overlook comprehensive 3D processing, potentially leading to a loss of critical spatial context. While promising results have been achieved using 2D slices of 3D imaging data, further research is needed to explore architectures that effectively integrate both 2D and 3D representational features. Such combined architectures may offer a more complete understanding of spatial relationships within the data, leading to improved performance in tasks involving volumetric medical imaging.

4 Research Design & Methodology

As previously noted, both 3D CNNs and transformer architectures offer distinct advantages and drawbacks. Convolutional operations, with their localized receptive fields, excel at capturing details in non-complex images where contexts are fixed. However, for 3D medical images, the relevant context often spans the entire image. On the other hand, transformer architectures can address this with their self-attention mechanism, which computes the global context of an image. Yet, they tend to under perform compared to CNNs when working with limited datasets, as is often the case with medical images, and they come with higher computational costs. Building on these considerations, our objective is to develop a hybrid architecture that integrates both CNN and Vision Transformer models, specifically tailored for the classification of 3D medical images. With this aim, we can formally define our research questions as follows.

4.1 Research Questions

4.1.1 RQ1

Can 3DViTMedNet combine the strengths of transformers and convolutional neural networks (CNNs) and hence set new benchmarks for performance on the 3D Medical MNIST dataset?

4.1.2 RQ2

How does the integration of 3D and 2D feature representations in 3DViTMedNet impact the overall accuracy and interpretability of medical image classification?

4.1.3 RQ3

Can 3DViTMedNet effectively capture both local and global spatial features in volumetric medical data, and how does it compare to purely convolutional or transformer-based models in terms of computational efficiency and model complexity?

4.2 Overview of Research Methods

4.2.1 Slice-Based Volumetric Image Processing techniques

As reviewed in the literature, transformer architectures for processing 3D input volumes often utilize 2D slices. This approach can be divided into two main methods:

Iterative Slicing: Several studies, such as [65], [22] and [29], iterate through the depth of the volume, processing each slice individually. While this method yields reasonable results, it poses a challenge in fields like medical imaging, where maintaining spatial context is crucial. Splitting a volume into slices and processing them independently can lead to loss of spatial information. Despite being less computationally intensive, this approach sacrifices the spatial context necessary for accurate medical image interpretation.

Multiplanar Slicing: Other studies, such as [33] and [35], process 3D volumes using slices from three different planes: coronal, sagittal, and axial. This method aims to preserve spatial information while still processing slices iterativly. By incorporating multiple planes, these models maintain a better understanding of the spatial context.

Computational Complexity:

The primary challenge in directly processing 3D images with transformer architectures stems from their high computational cost, particularly due to the quadratic complexity of vision transformers [19]. As a result, processing volumetric data for medical image classification can be prohibitively expensive. Interestingly, there is a notable absence of models designed to handle volumetric inputs directly for classification tasks, highlighting a trade-off between computational efficiency and model performance that warrants further exploration.

In the context of medical imaging, patient outcomes take precedence. Therefore, if increased computational demands can lead to significantly improved model performance, this trade-off becomes justified and motivates our research inquiries.

In practice, there are several compelling reasons for processing 3D images as 2D slices, including (but not limited to) the following

Limited Data Availability:

Medical datasets often contain fewer volumetric 3D images compared to 2D images. Training 3D CNNs requires large amounts of data to avoid over fitting and ensure model generalization. When data is limited, slicing 3D volumes into 2D images increases the effective dataset size, enabling better training with the available data.

Pre trained Model Availability:

There are many pre trained models available for 2D CNN architectures, which can be fine-tuned for medical image classification. These pre trained models are often trained on large 2D image datasets (like ImageNet), and transferring this knowledge is easier with 2D slices than with 3D images, where pre trained models are scarce.

Memory Efficiency:

3D volumes consume significantly more memory than 2D slices. Training 3D CNNs requires large amounts of GPU memory, which can be a limiting factor, especially when working with high-resolution medical images. By using 2D slices, the memory requirements are drastically reduced, allowing for more efficient training.

4.2.2 Developing a hybrid architecture

With a clear motivation for our approach to processing input data, we can now outline the development of our architecture. While 3D CNN architectures have been the standard for medical imaging tasks due to their established efficacy, they are limited by the receptive field of convolutional operations. This limitation can result in suboptimal performance for certain computer vision tasks. Transformer models, on the other hand, overcome this by leveraging the attention mechanism, which inherently allows for a larger receptive field.

However, the scarcity of large datasets in the 3D medical imaging space poses a challenge for pure transformer architectures. These models often struggle to learn meaningful representations compared to CNNs, which can do so with relatively minimal data. This issue is underscored in Google's ViT paper [19], which highlights the need for over 300 million private images to achieve state-of-the-art classification accuracy.

Given the advantages and limitations of both transformer and CNN architectures, a hybrid approach that combines the strengths of both is particularly promising for medical imaging. This approach can effectively capture both local and global context within an image, which is crucial for identifying critical information dispersed throughout the volume.

The proposed research adopts a hybrid architecture designed to optimize both local and global feature extraction. Initially, volumetric data is input into a 3D CNN to capture local context and down-sample the images. The resulting output is then divided into slices across each dimension and passed through a pre trained 2D CNN, which extracts representational features from each plane. Instead of relying on a traditional fully connected layer for classification, we introduce a transformer architecture to capture the global context across the entire volume. The final classification is produced by a MLP.

To address the challenge of generalization on smaller datasets—particularly in transformer models—our architecture incorporates a pre trained CNN along with a robust data augmentation pipeline. This augmentation strategy aims to increase the diversity of the input data, enhancing the model's overall performance. We anticipate that the evaluation metrics will validate the computational investment, further proving the model's efficacy.

By combining the CNN's ability to extract detailed local features with the transformer's strength in modeling global relationships, this hybrid approach aims to significantly improve performance

in medical image classification tasks.

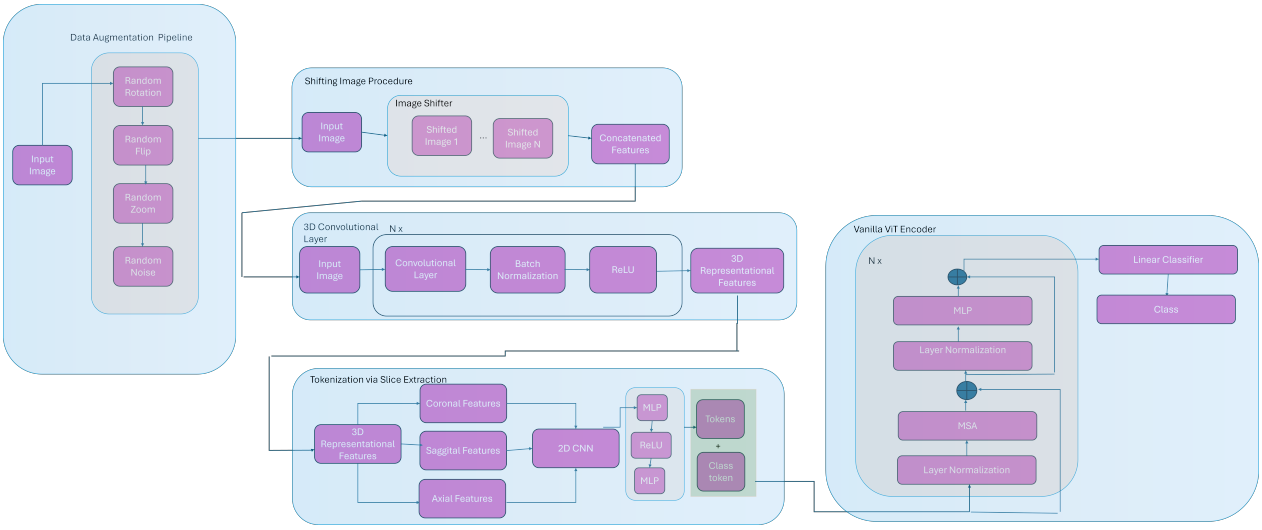

Figure 21: Visualization of the proposed high level architecture

4.2.3 Data Augmentation Pipeline

As previously mentioned, the integration of a data augmentation pipeline offers a means to augment the training dataset, a crucial step, particularly in the realm of 3D medical imaging where data scarcity is prevalent. Illustrated in the data augmentation pipeline, our proposed data augmentation pipeline delineates each augmentation technique, provided in the following enumeration: **Data Augmentation Pipeline for 3D Image:**

1. **Resize:** Let I denote the original 3D image with dimensions $D \times H \times W$ (depth $\times$ height $\times$ width). After resizing, the new dimensions become $D' \times H' \times W'$.

2. **Normalize:** For each voxel v in the resized image, normalize its intensity value using:

$$v_{\text{normalized}} = \frac{v - \min(I)}{\max(I) - \min(I)}$$

3. **Rotate:** Rotate the 3D image by an angle θ (in degrees) along the specified axes (e.g., x, y, or z axis). The rotation can be represented using rotation matrices specific to 3D transformations. For a rotation around the x-axis:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Similar transformations can be derived for rotations around the y and z axes.

4. **Flip:** Flip the 3D image along the specified axes (e.g., horizontally, vertically, or depth-wise). For each axis, reverse the order of voxels along that axis.

5. **Zoom:** Zoom into or out of the 3D image by a scale factor s along each dimension (D, H, and W). This can be achieved by resizing the image along each dimension using the formula:

$$\text{zoomed\_depth} = s \times D, \quad \text{zoomed\_height} = s \times H, \quad \text{zoomed\_width} = s \times W$$

6. **Add Noise:** Add Gaussian noise to each voxel in the 3D image. For each voxel v, add a random value sampled from a Gaussian distribution with mean μ and standard deviation σ:

$$v_{\text{noisy}} = v + \epsilon, \quad \epsilon \sim \mathcal{N}(\mu, \sigma^2)$$

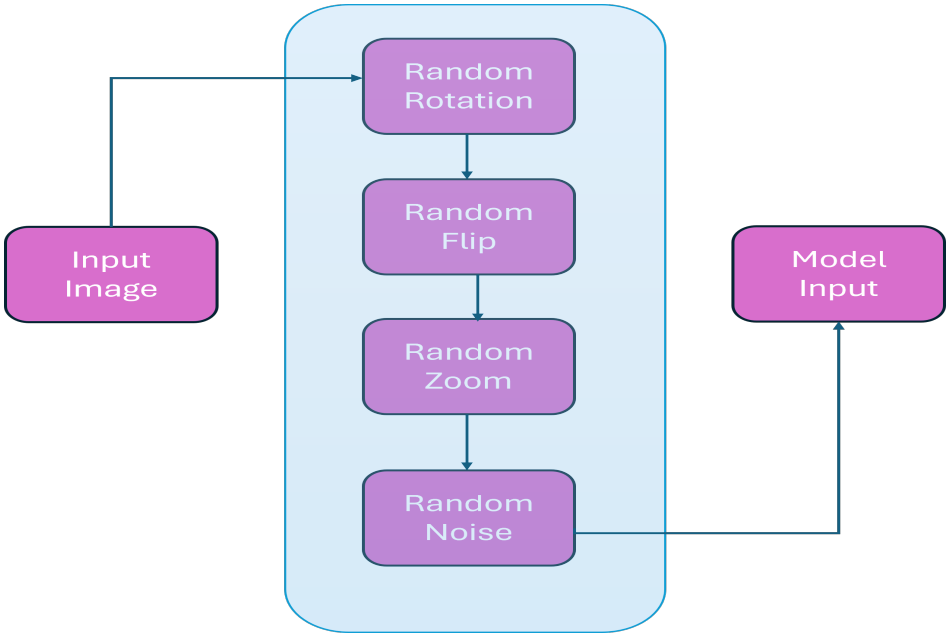

Figure 22: Visualization of the data augmentation pipeline

4.2.4 Image Shifting Policy

In the ViT architecture, an image is divided into non-overlapping patches, which are then flattened and embedded into patch tokens. While this reduces the computational complexity and allows transformers to handle image data, it struggles with capturing local continuity and fine-grained spatial features between patches.

To address this, we employ an image shifting policy introduces overlapping patches. By shifting the input images slightly in different directions (e.g., up, down, left, right), the model can capture more local details and preserve spatial context more effectively. This provides a richer representation of the input image.

Given a 2D image $I \in \mathbb{R}^{H \times W \times C}$, where H, W, and C are the height, width, and channels of the image, respectively,we can create extracts non-overlapping sub-images of size $P \times P$. This can be described mathematically as:

$$I_p = \{I_{i,j} \mid 0 \leq i < \frac{H}{P}, 0 \leq j < \frac{W}{P}\}$$

where I_p represents the set of non-overlapping images. In our image shifting policy, a slight shift is applied along both the x and y axes, resulting in overlapping patches.

We shift the patch window by an offset s in each direction, yielding overlapping patches. This can be represented as:

$$I_{shifted} = \{I_{i+s,j+s} \mid 0 \leq i < \frac{H}{P}, 0 \leq j < \frac{W}{P}\}$$

The overlapping Images are then passed through the 3D CNN 23.

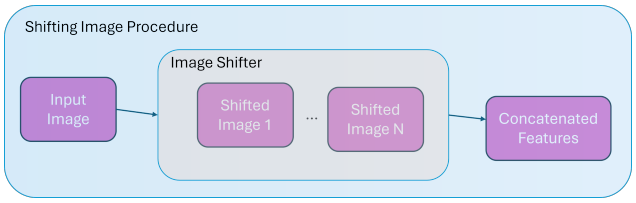

Figure 23: Visualization of the Image Shift Policy

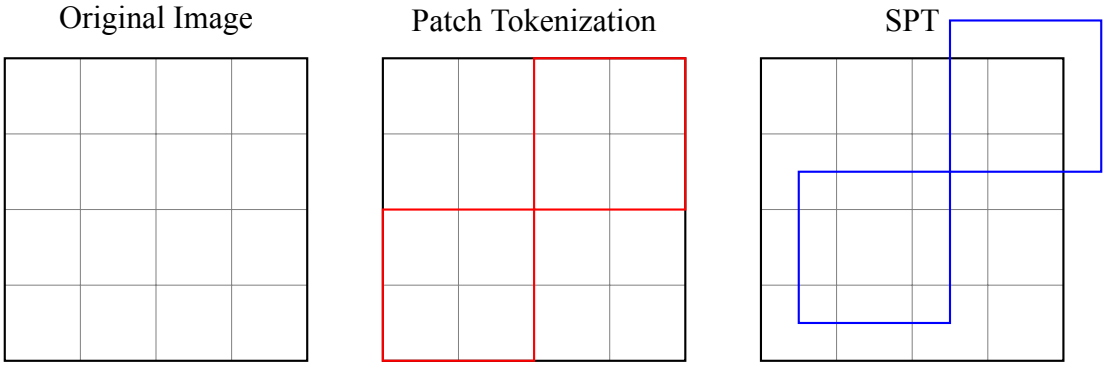

Figure 24: Comparison of Standard vs Shifted Patch Tokenization (SPT)

4.2.5 3D CNN Design

To generate 3D representational features of the input volume, we propose using a $3 \times 3 \times 3$ convolutional layer followed by batch normalization and a ReLU activation function. This combination is particularly effective for capturing local spatial features within the 3D input, as the $3 \times 3 \times 3$ convolutional kernel can learn to recognize patterns and structures that are

three-dimensional in nature. Batch normalization helps to stabilize and accelerate the training process by normalizing the output of the convolutional layer, while the ReLU activation function introduces non-linearity, allowing the network to model more complex functions.

Repeating this process with another $3 \times 3 \times 3$ convolutional layer, followed again by batch normalization and ReLU, further refines these features and enhances the network's ability to capture intricate spatial relationships within the data. By stacking these layers, the network can progressively learn higher-level abstractions and more detailed representations of the input volume. This deep feature extraction process is crucial for effectively understanding and processing 3D data, setting a strong foundation for subsequent stages of the neural network, such as feeding these features into a 3D transformer model for advanced analysis and decision-making tasks.

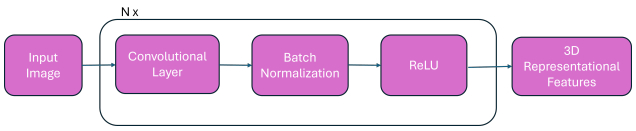

Figure 25: Visualization of the proposed 3D CNN

Hence we can represent the 3D representational features X generated from input J as

$$X = \mathbf{D}_{\text{3d\_representation}}(J)$$

where

$$\mathbf{D}_{\text{3d\_representation}}$$

represents a singular 3D CNN block described in Figure 22.

4.2.6 2D Pre trained CNN

The 2D CNN receives inputs from the Coronal, Sagittal, and Axial planes, derived from the 3D representational features. This model is an implementation of the ResNet-50 architecture, pre trained on the ImageNet database. The pre trained weights are further fine-tuned using corresponding 2D image datasets from the same volumetric image source, enhancing the model's ability to capture relevant features for each plane.

4.2.7 Projection Layer

From here we must generate tokens from our combined representational features, here we incorporate a vanilla non linear projection layer.

Given the recently generated 3D Representational features $\mathbf{X} \in \mathbb{R}^{D \times H \times W \times C}$, where H, W, and C represent the height, width, and number of channels, respectively, we perform the following steps:

1. **First MLP Layer**: The first MLP layer transforms the input channel dimension C to an intermediate dimension d_{h}:
$$\mathbf{Y} = \text{ReLU}(\mathbf{W}_1 \mathbf{X} + \mathbf{b}_1)$$
where $\mathbf{W}_1 \in \mathbb{R}^{C \times d_{\text{h}}}$ and $\mathbf{b}_1 \in \mathbb{R}^{d_{\text{h}}}$ are the weights and biases of the first MLP layer.
2. **ReLU Activation**: The ReLU activation function introduces non-linearity:
$$\mathbf{Y}_{\text{ReLU}} = \max(0, \mathbf{Y})$$
3. **Second MLP Layer**: The second MLP layer transforms the intermediate dimension d_{h} to the final projection dimension d:
$$\mathbf{T} = \mathbf{W}_2 \mathbf{Y}_{\text{ReLU}} + \mathbf{b}_2$$
where $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{h}} \times d}$ and $\mathbf{b}_2 \in \mathbb{R}^{d}$ are the weights and biases of the second MLP layer.
4. **Final Output**: The final output $\mathbf{T} \in \mathbb{R}^{D \times H \times W \times d}$ is the result of the non-linear projection.

The projection process can be summarized as:

$$\mathbf{T} = \mathbf{D}_{\text{mlp}}(\mathbf{X}) = \mathbf{W}_2 \max(0, \mathbf{W}_1 \mathbf{X} + \mathbf{b}_1) + \mathbf{b}_2$$

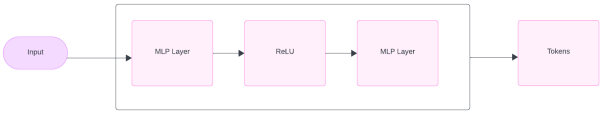

Figure 26: Visualization of the proposed projection Layer

4.2.8 Transformer Design

The ViT implementation follows the standard approach used in most transformer-based architectures. As outlined in the background, the initial step involves projecting our input into an embedding space and applying positional encoding to each generated patch, as transformers require sequential input. These patches are then fed into the transformer encoder. The output from the transformer encoder is subsequently processed by a basic MLP layer, which generates the classification results.

The implementation details of the transformer encoder block are crucial to the transformer architecture. From the literature, it is common to use either the Vision Transformer (ViT) architecture or the Swin Transformer architecture. Both have their advantages and disadvantages. Swin Transformers have been shown to reduce computational complexity but often fall short in achieving higher classification accuracy. Conversely, ViT achieves superior classification accuracy but at a higher computational cost.

Notably, while there are numerous studies comparing ViT and Swin Transformers for 2D images, there is a lack of justification for choosing one over the other in 3D medical image classification. The debate typically centers around computational cost rather than classification accuracy.

Given our objective to classify medical images with the highest accuracy possible, our initial (V0) architecture will incorporate a 3D implementation of ViT.

With the proposed architecture in mind, we will be able to successfully develop a hybrid architecture which can classify the MedMNIST 3D Datasets. Through generating evaluation and efficiency metrics specified below we will be able to compare and contrast benchmark accuracy's and area under curves with out proposed model.

5 Research Plan

The proposed research above will abide by a SCRUM framework. We note that as we already have access to the 3D Medical MNIST Dataset (V1 and V2 respectively) [63], [64] which is already normalised. With this proposed Sprint design in figure 9, we plan to perform 2 week sprints until we gain sufficient results (these results will be documented further in the 'Evaluation Metrics' Section of this document). Once we have achieved satisfactory results (which will be determinate via stakeholder agreement), we look to expand into optimizing the models efficiency followed by formalizing results in a report.

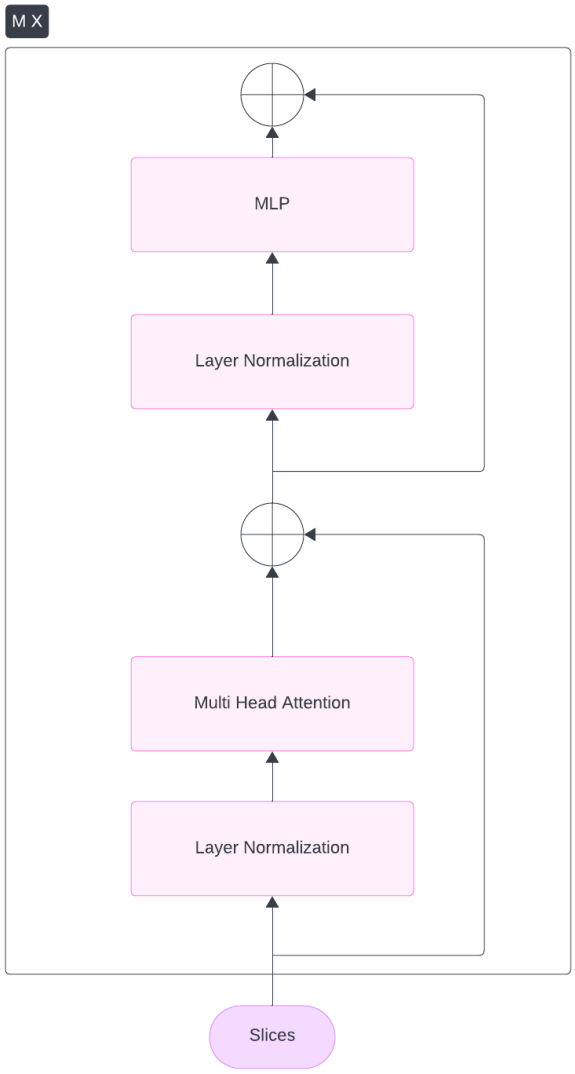

Figure 27: Visualization of the proposed ViT

Below, we outline key milestones for the project, along with their respective deadlines that have either been met or are anticipated to be met. This research plan helps us track the progress of the project. It is important to note that, at the time of submission, our current progress does not include the implementation of hyper-parameter optimization or a data augmentation pipeline. We anticipate that incorporating these elements will enable our model to surpass the benchmark models effectiveness metrics on most of the datasets.

- Codifying the proposed architecture, excluding data augmentation pipeline and hyper-parameter optimization techniques. (Completed 15/04/24)
- First MVP deployment (Completed 17/04/24)
- First MVP evaluation/comparison (Completed 25/04/24)
- First MVP deployment via Deakin Universities HPC (Completed 15/05/24)

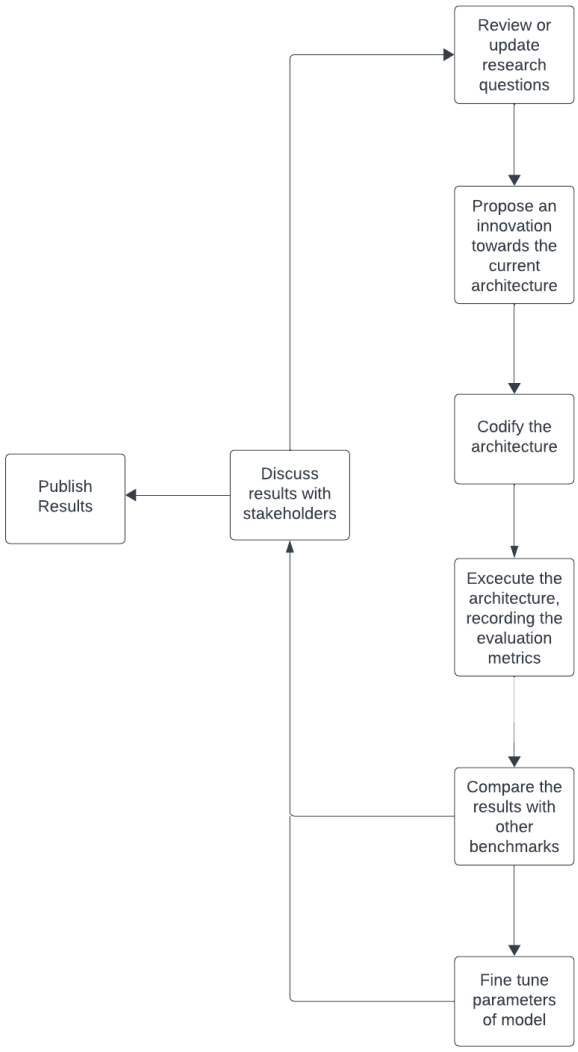

Figure 28: Visualization of the SCRUM frameworks iterative sprint design

- Submitting Research proposal (Completed 31/05/24)
- Research Defence (Completed 14/06/24)
- Deploying our optimized MVP (Completed 12/08/24)
- Optimized MVP comparison (Completed 13/08/24)
- Training model on Deakin GPU Cluster (Completed 13/08/24)
- Redesigning model to improve local feature attainment (Completed 25/08/24)
- Incorporating Shifted Patch Tokenization (Completed 30/08/24)
- Integrating the Data- Augmentation Pipeline (Completed 05/09/24)
- Deploying our final MVP prior to conference submissions (Completed 15/09/24)
- Final evaluation of MVP with benchmark models (Completed 20/09/24)

- Codifying other transformer architectures for comparison(Completed 20/09/24)
- Incorporating our final model into final research submission (Completed 07/10/24)
- Research Defence (Expected 13/10/24)

5.1 Sustainability

Long-term Impact and Relevance

To ensure this research project makes a lasting impact in the field of 3D medical image classification, it is essential not only to achieve competitive or state-of-the-art results but also to explore novel methods that have yet to be applied in hybrid transformer architectures. Additionally, the proposed architecture must be evaluated for its generalizability across various 3D medical imaging datasets. For instance, achieving strong results solely on the BREASTMNIST dataset [63], while valuable, may limit the broader impact of the research. Broader applicability across multiple datasets would significantly enhance its long-term relevance and contribution to the field.

Dissemination of Findings

The findings of this project will be disseminated through multiple channels to reach a diverse audience. Research outcomes are anticipated to be presented at conferences in late 2024 or 2025, facilitating engagement with scholars and professionals in the field. Additionally, platforms such as Medium and YouTube will be utilized to extend the reach of the research to a broader audience, including both experts and non-experts interested in 3D medical image classification techniques. This comprehensive dissemination strategy ensures that the research findings are accessible and impactful across various communities.

Data accessibility

For stakeholders interested in the results and code base generated from this experiment, a public repository will be maintained and openly accessible. This will allow stakeholders to execute the model, compare results, and explore the findings in greater detail. Comprehensive documentation will be provided in the repository's README section, detailing the steps required to reproduce the results obtained from the experiment. Additionally, the code base will be thoroughly commented to ensure clarity and ease of understanding.

Promoting sustainability of research outcomes

Funding can be sought after to increase the scope of this research to produce more insightful results and outcomes. Furthermore, medical (or other) institutions can collaborate with key

stakeholders in this project to create policies and practices with these aforementioned institutions for their own deep learning classification pipelines.

6 Experimental Setup

6.1 Datasets

In selecting datasets for this research, it was crucial that they were medical in nature to ensure relevance to the field of 3D image classification and to address specific medical imaging challenges. The use of medical datasets, particularly those containing imaging data, provides an ideal scenario for assessing the capabilities of transformer-based models in processing complex, high-dimensional data typical of the medical domain. These datasets must reflect real-world medical scenarios, such as those involving disease diagnosis or anatomical classification. They offer a variety of challenges such as noise, variability in image acquisition, and a high degree of interpretive difficulty, which are key factors that can influence model performance in medical applications.

Additionally, the datasets needed to contain a variety of imaging modalities (e.g., MRI, CT, and PET scans) to ensure generalizability of the models across different types of 3D medical images. This diversity in modalities allows for a thorough evaluation of the model's ability to extract meaningful features regardless of imaging technique. It was also essential that the datasets exhibit consistency between features in images (e.g., standardized resolutions and aligned anatomical structures) to prevent bias and allow for meaningful comparison between images. Furthermore, due to the time constraints associated with an honours project, only pre-curated datasets were selected. These datasets, which are already cleaned and labeled, enable rapid experimentation without the significant time investment required for raw data pre processing. We applied 3DViTMedNet to six datasets selected from the database outlined below, yielding varying degrees of success across the different datasets.

6.1.1 MedMNIST

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 29, 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 [63].

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

Figure 29: The MedMNIST3D dataset 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).

The following provides a comprehensive explanation of each dataset utilized in the experimental phase.

OrganMNIST3D is derived from the Liver Tumor Segmentation Benchmark (LiTS) [11]. This dataset is designed for multi-class classification across 11 body organs, comprising a total of 1,742 samples 30.

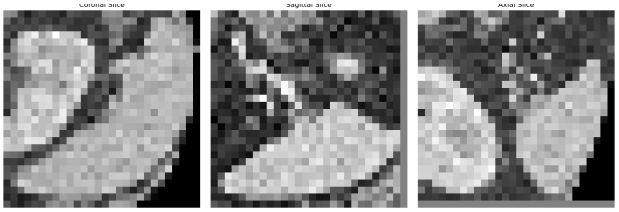

Figure 30: Visualization of the coronal, sagittal, and axial planes from the OrganMNIST3D dataset in the MedMNIST collection.

NoduleMNIST3D is derived from the LIDC-IDRI31 dataset [8], 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 31.

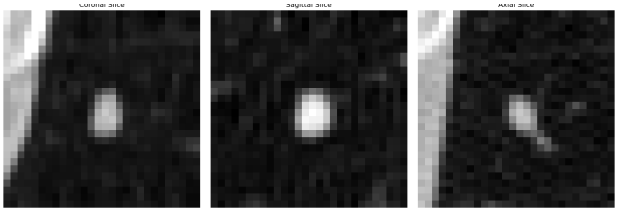

Figure 31: Visualization of the coronal, sagittal, and axial planes from the NoduleMNIST3D dataset in the MedMNIST collection.

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 32.

Figure 32: Visualization of the coronal, sagittal, and axial planes from the AdrenalMNIST3D dataset in the MedMNIST collection.

FractureMNIST3D is derived from the RibFrac Dataset [34], 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. 33

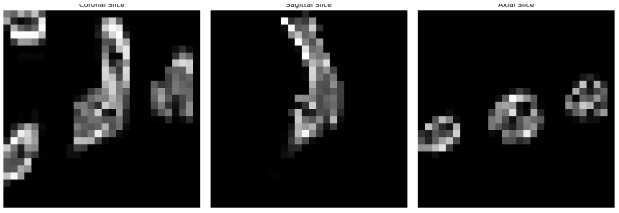

Figure 33: Visualization of the coronal, sagittal, and axial planes from the FractureMNIST3D dataset in the MedMNIST collection.

VesselMNIST3D is based on the IntrA33 dataset [46], 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 34.

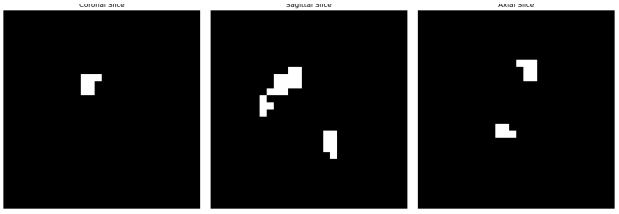

Figure 34: Visualization of the coronal, sagittal, and axial planes from the VesselMNIST3D dataset in the MedMNIST collection.

SynapseMNIST3D is a 3D volume dataset based on the MitoEM dataset [60] 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 35.

To ensure the reliability and validity of the data captured from the MedMNIST database, several key measures were implemented throughout the research process. The MedMNIST database is well-curated, having undergone preprocessing steps such as normalization and resizing, which enhances the consistency and comparability of the images. This preprocessing minimizes the risk of errors stemming from variations in image quality or size. Additionally, the datasets provided in

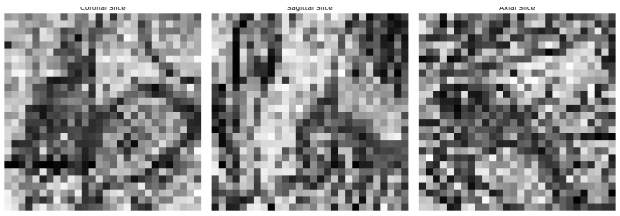

Figure 35: Visualization of the coronal, sagittal, and axial planes from the SynapseMNIST3D dataset in the MedMNIST collection.

MedMNIST are labeled and organized based on standardized protocols, ensuring that the labels are accurate and the dataset is representative of real-world medical scenarios.

Potential biases were addressed by utilizing a diverse set of datasets within MedMNIST, encompassing different imaging modalities (e.g., MRI, CT) and medical conditions. This diversity reduces the risk of overfitting the model to a specific type of medical image, thus improving the generalizability of the results. Furthermore, cross-validation techniques were employed during model training to minimize the influence of data splits on the outcomes.
To address any remaining limitations, we acknowledged inherent challenges, such as class imbalances or dataset-specific biases, and took steps to mitigate these by employing techniques like data augmentation and class weighting. This combination of strategies helps to ensure the data is both reliable and valid for use in the classification tasks.

A notable limitation of this research is that there has not been access to any medical professionals, and as such, the accuracy of the study relies on the correctness of the labels provided in the MedMNIST database. While MedMNIST is a well-curated and widely accepted resource, the datasets within it are sourced from multiple independent providers, each with its own curation process. This structure introduces a degree of uncertainty, as the research assumes that the labeling and data quality across all datasets are accurate. It is important to note, however, that if a particular dataset were to contain any issues, this would not imply that the remaining datasets are similarly affected. Each dataset is curated independently, minimizing the risk that problems in one dataset would propagate to others, thereby maintaining the integrity of the research across the diverse datasets used.

6.2 Implementation Details

6.2.1 Hyperparameters

In the implementation of our 3DViTMedNet architecture, several key hyperparameters were carefully selected to optimize model performance. We utilized the Adam optimizer with a learning rate of 1×10^{-5}. Additionally, a MultiStepLR learning rate scheduler was employed with specified milestones and a decay factor γ , allowing the learning rate to decrease at predefined

stages, which helps stabilize the learning process over the course of training. The model was trained for 100 epochs, with a batch size of 6 per iteration.

For the 3D CNN component of the architecture, a kernel size of 5 was applied, along with a stride of 1 and padding of 2, ensuring that the convolutions adequately capture spatial features while maintaining dimensional consistency. To extract features from different views, we stacked 3 slices from each dimension—coronal, sagittal, and axial—resulting in the corresponding planes having three times as many channels for the input to the 3D CNN.

The transformer module of the network utilized 8 attention heads per layer, striking a balance between computational complexity and the ability to capture long-range dependencies in the data. The transformer layers process the tokenized representations of the 3D image features, enabling the model to focus on critical aspects of the data. For binary classification tasks, such as distinguishing between Alzheimer's Disease (AD) and normal controls (NC), binary cross-entropy loss was used as the loss function.

6.2.2 Software and Hardware requirements

The implementation was carried out using the Python interface of TensorFlow [reference]. TensorFlow provides a comprehensive set of pre-built neural network functionalities, facilitates automatic differentiation, and offers a robust environment optimized for deep learning applications. Additionally, the MedMNIST [referehce] library was employed to streamline the collection and management of dataloaders, significantly simplifying the training and testing processes. For visualization and data preprocessing tasks, libraries such as Matplotlib [reference] and NumPy [reference] were utilized, ensuring efficient handling of image data and result interpretation.

The model was trained on NVIDIA A100 Tensor Core GPUs, leveraging their computational power to handle the complexity of the architecture and the large datasets. Total training time across all datasets was approximately 25 hours, allowing for the completion of multiple epochs while maintaining a reasonable training duration.

This combination of hyperparameters and computational resources ensured that our model was able to effectively learn from the data, yielding optimal results in the classification tasks.

7 Empirical Evaluation

7.1 Overview of Performance Evaluation and Statistical Validation

For each dataset, we conducted a thorough evaluation by assessing class distribution and outlier detection to ensure data integrity and to identify potential biases. To evaluate the performance of our 3DViTMedNet model, we utilized key classification metrics, specifically accuracy and the area under the curve (AUC), which allowed us to benchmark our model against the quantitative standards set by the baseline models. These metrics provide a comprehensive understanding of the model's ability to distinguish between classes, particularly in medical image classification tasks.

To further enhance the reliability of our results, we employed k-fold cross-validation, which partitions the dataset into multiple folds and iterates the training process over different folds. This technique mitigates the risk of overfitting by ensuring that the model is tested on various subsets of the data, resulting in a more robust performance estimate. Additionally, to confirm the statistical significance of performance differences, we applied the Wilcoxon signed-rank test, a non-parametric statistical test that compares paired results to determine if the observed differences are meaningful.

To gain deeper insights into model behavior, we also leveraged attention map visualizations to interpret which regions of the images the model focused on during classification. This helped in understanding model decision-making processes. Moreover, loss visualizations were employed to track training and validation loss over time, aiding in the analysis of model convergence and potential overfitting during training. To understand model complexity we also compare floating point operations (FLOPS) between models.

7.2 OrganMNIST3D Dataset: Classification Metrics and Model Performance

The classification performance on the OrganMNIST3D dataset, as shown in Table 1, highlights the efficacy of different models in medical image classification tasks. The ResNet-18 3D model stands out by achieving the highest AUC (0.996) and accuracy (0.907), which underscores the model's ability to capture both local and global features effectively through its 3D convolutional structure. This result demonstrates that the ResNet-18 3D model, which benefits from a deeper network and 3D convolutions, is particularly well-suited for recognizing intricate anatomical features present in the OrganMNIST3D dataset.

In comparison, our proposed 3DViTMedNet achieves competitive results, with an AUC of 0.963 and accuracy of 0.726. These results reflect the model's ability to blend the strengths of both 3D CNNs for local feature extraction and Vision Transformers (ViT) for capturing global

dependencies.

OrganMNIST3D Classification Metrics		
Methods	AUC	ACC
ResNet-18 + 2.5D	0.977	0.788
ResNet-18 + 3D	<u>0.996</u>	<u>0.907</u>
ResNet-18 + ACS	0.994	0.900
ResNet-50 + 2.5D	0.974	0.769
ResNet-50 + 3D	0.994	0.883
ResNet-50 + ACS	0.994	0.889
auto-sklearn	0.977	0.814
AutoKeras	0.979	0.804
3D ViT	0.636	0.325
3D SWIN	0.956	0.673
3DVITMEDNET (Ours)	0.963	0.726

Table 1: Classification performance of models on the OrganMNIST3D dataset

The attention maps generated by the 3DViTMedNet 37 exhibit a more global focus, attempting to capture broad, distributed patterns across the image slices. In the coronal, sagittal, and axial planes, the attention appears spread across multiple key positions, indicating that the transformer is analyzing relationships across different areas of the image simultaneously. This is consistent with the goal of transformer models, which are designed to capture long-range dependencies and global context within the input. However, this broad attention could be less effective in identifying fine-grained, local structures that are crucial in medical imaging tasks.

In contrast, the attention maps generated by the ResNet architecture exhibit a much more localized focus 36. The ResNet maps show distinct regions of high attention concentrated in smaller, specific areas. This suggests that the ResNet model is better at capturing local patterns and features within each slice, which aligns with the convolutional nature of ResNet models. These models are designed to capture hierarchical features through local convolutions, making them effective at identifying localized structures in medical images, such as tumors or lesions.

7.3 NoduleMNIST3D Dataset: Classification Metrics and Model Performance

The results for the NoduleMNIST3D dataset suggest that while various models achieved competitive performance, our proposed 3DViTMedNet model may need further review and optimization 2.

ResNet models performed consistently well, with ResNet-50 + ACS achieving the highest AUC (0.886) and accuracy (0.889), demonstrating the model's capability to capture local and hierarchical features in the data effectively. AutoML methods, such as auto-sklearn and

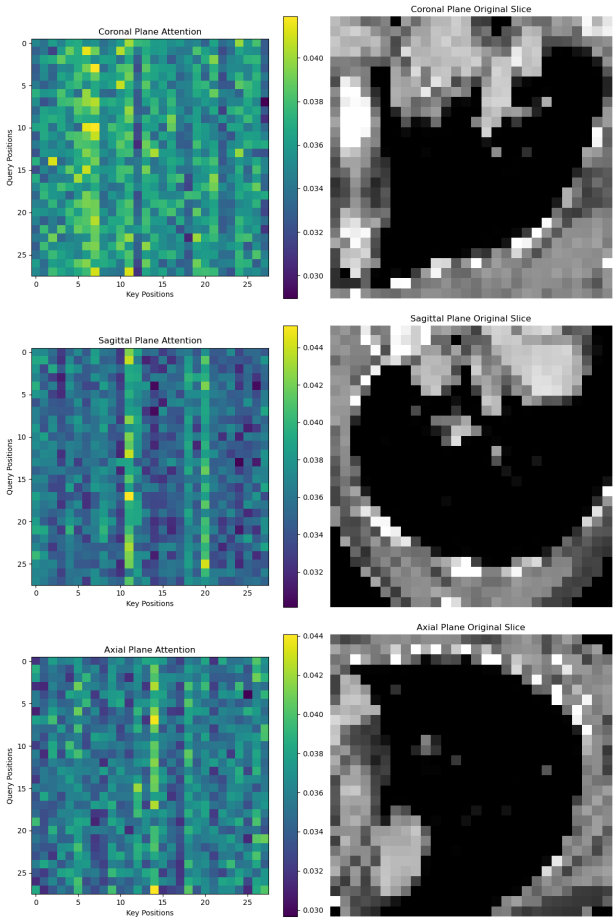

Figure 36: Visualization of the attention captured from the coronal, sagittal, and axial planes from the OrganMNIST3D dataset in the 3DVITMEDNET model.

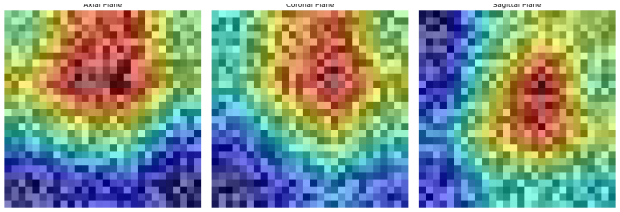

Figure 37: Visualization of the attention captured from the coronal, sagittal, and axial planes from the OrganMNIST3D dataset in the ResNet-18 + 3D.

AutoKeras, also performed well, with auto-sklearn achieving the highest overall AUC (0.914) and accuracy (0.874), likely due to their automatic optimization of hyperparameters and model selection. Transformer-based models (3D ViT and 3D SWIN) did not perform as well as ResNet models, which suggests that these architectures may struggle with the NoduleMNIST3D dataset, likely due to the need for stronger local feature extraction. 3DViTMedNet, in particular, achieved an AUC of 0.774 and accuracy of 0.794, indicating that the model may not be effectively capturing the critical features necessary for accurate classification in this dataset.

This is also evident from our qualitative analysis of the attention maps. While 3DViTMedNet does show increased attention in regions corresponding to the presence of nodules 38, it

struggles to consistently maintain its focus on these critical areas. In contrast, the ResNet-50 model 39 demonstrates a much more concentrated and persistent attention towards the nodule regions. This suggests that the ResNet-50 model has a better capacity for capturing and retaining the relationships between important local features in the NoduleMNIST3D dataset. The high attention displayed by the ResNet model likely contributes to its superior classification performance, as it can more effectively learn the local anatomical structures essential for accurate diagnosis. This disparity highlights the need for further refinement of 3DViTMedNet in order to improve its focus and feature extraction capabilities in complex medical image classification tasks.

NoduleMNIST3D Classification Metrics		
Methods	AUC	ACC
ResNet-18 + 2.5D	0.838	0.835
ResNet-18 + 3D	0.863	0.844
ResNet-18 + ACS	0.873	0.847
ResNet-50 + 2.5D	0.835	0.848
ResNet-50 + 3D	0.875	0.847
ResNet-50 + ACS	0.886	0.889
auto-sklearn	<u>0.914</u>	<u>0.874</u>
AutoKeras	0.844	0.834
3D ViT	0.795	0.79
3D SWIN	0.80875	0.83226
3DVITMEDNET (Ours)	0.774	0.794

Table 2: Classification performance of models on the NoduleMNIST3D dataset

7.4 Adrenal3DMNIST Dataset: Classification Metrics and Model Performance

The results for the AdrenalMNIST3D dataset indicate a strong performance for our proposed 3DViTMedNet model, which achieves the highest AUC (0.883) and accuracy (0.846) among all models tested 3. These results are statistically significant, as they outperform other architectures by a considerable margin, particularly in AUC, which reflects the model's ability to discriminate between classes.

In comparison, the ResNet models, while also performing well, did not surpass 3DViTMedNet in terms of both AUC and accuracy. For instance, ResNet-50 + ACS obtained an AUC of 0.828 and accuracy of 0.758, showcasing competitive results but falling short of the performance of 3DViTMedNet. Similarly, auto-sklearn displayed good accuracy (0.802), but its AUC of 0.828 indicates that its classification performance does not match that of our model.

Transformer-based models such as 3D ViT and 3D SWIN struggled on this dataset, with AUC values of 0.549 and 0.708, respectively. These results further emphasize the significance of

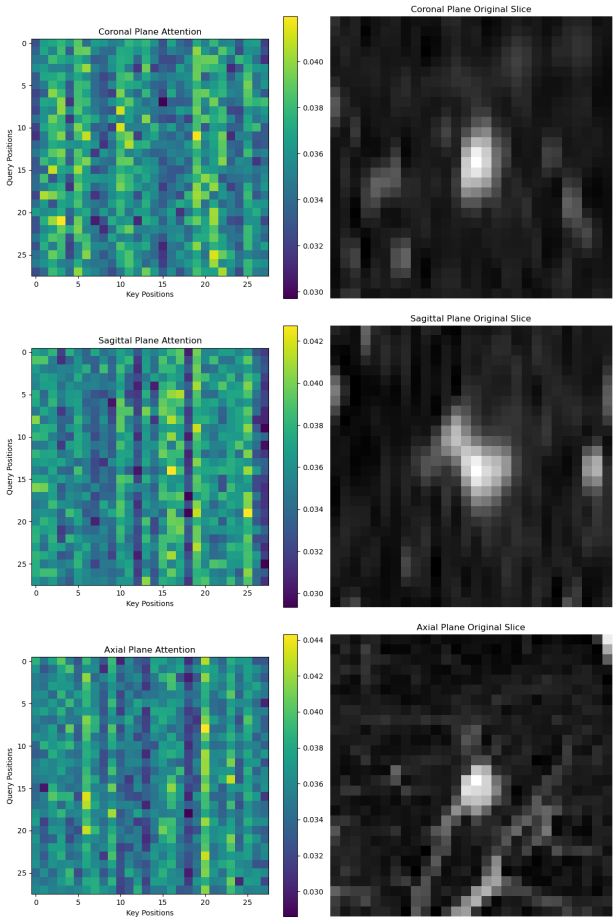

Figure 38: Visualization of the attention captured from the coronal, sagittal, and axial planes from the NoduleMNIST3D dataset in the 3DVITMEDNET model.

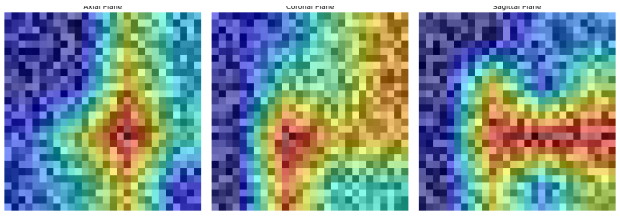

Figure 39: Visualization of the attention captured from the coronal, sagittal, and axial planes from the NoduleMNIST3D dataset in the ResNet-50 + ACS.

3DViTMedNet's architecture in capturing both local and global relationships within the data, leading to superior classification performance. Overall, the results indicate that 3DViTMedNet is highly effective for the AdrenalMNIST3D dataset, with statistically significant improvements over other models.

Interestingly, when we compare the 3DViTMedNet attention map 40 with the ResNet attention map 41, we observe a notable difference in behavior for the AdrenalMNIST3D dataset compared to previous datasets. Unlike prior results, 3DViTMedNet displays a heightened focus towards the center of the scan, suggesting that the model has been able to establish localized relationships more effectively. This could be attributed to the nature of the scans within this specific dataset,

allowing 3DViTMedNet to identify and attend to relevant regions more quickly than in previous datasets. The ability to form such localized relationships may explain its superior performance on this dataset, highlighting the adaptive nature of the model when confronted with varying medical imaging contexts.

AdrenalMNIST3D Classification Metrics		
Methods	AUC	ACC
ResNet-18 + 2.5D	0.718	0.772
ResNet-18 + 3D	0.827	0.721
ResNet-18 + ACS	0.839	0.754
ResNet-50 + 2.5D	0.732	0.763
ResNet-50 + 3D	0.828	0.745
ResNet-50 + ACS	0.828	0.758
auto-sklearn	0.828	0.802
AutoKeras	0.804	0.705
3D ViT	0.549	0.769
3D SWIN	0.708	0.769
3DVITMEDNET (Ours)	<u>0.883</u>	<u>0.846</u>

Table 3: Classification performance of models on the AdrenalMNIST3D dataset

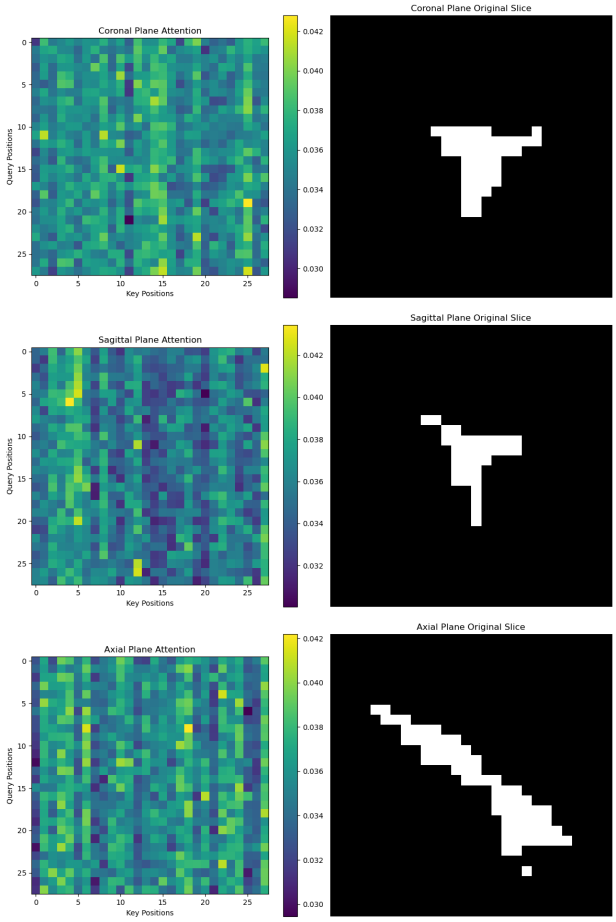

Figure 40: Visualization of the attention captured from the coronal, sagittal, and axial planes from the AdrenalMNIST3D dataset in the 3DVITMEDNET model.

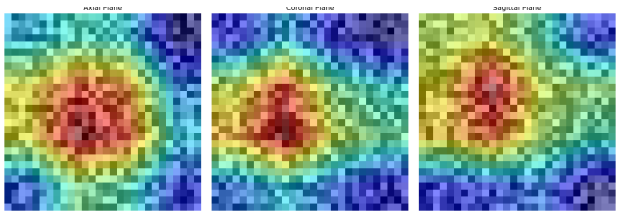

Figure 41: Visualization of the attention captured from the coronal, sagittal, and axial planes from the AdrenalMNIST3D dataset in the ResNet-18 + ACS.

7.5 FractureMNIST3D Dataset: Classification Metrics and Model Performance

The results on the FractureMNIST3D dataset indicate that performance across all models is generally lower compared to other datasets, with no model achieving statistically significant AUC or accuracy scores 4. ResNet-50 + ACS achieves the best performance with an AUC of 0.750 and an accuracy of 0.517, suggesting that this configuration of ResNet was able to capture relevant features more effectively than others.

Interestingly, both auto-sklearn and AutoKeras exhibit moderate performance with AUC scores around 0.628 and 0.642, respectively, but their accuracy remains low, indicating potential difficulty in learning from the dataset. The 3DViTMedNet model achieved an AUC of 0.631 and an accuracy of 0.41, showing competitive results compared to some of the other models but still under performing compared to the top-performing ResNet-50 + ACS.

Overall, these results suggest that the FractureMNIST3D dataset poses a challenge to all models, particularly in terms of achieving higher accuracy, and further optimization of architectures or different approaches may be required to improve performance on this dataset. It also highlights that while transformer-based models like 3DViTMedNet are competitive, they may require additional tuning to capture critical features in this specific medical imaging task.

A potential reason for the suboptimal performance of all models on the FractureMNIST3D dataset could be attributed to the large amount of empty space within the images or the relatively small size of the region of interest (ROI) containing the fractures. This limitation may hinder the ability of the models to effectively identify and focus on the relevant features. As reflected in the attention map 42, 3DViTMedNet demonstrates increased attention in areas that are roughly aligned with the fracture, but it struggles to produce a refined and concentrated attention area. A similar issue is observed in the ResNet attention map 43, where the model's attention does not align precisely with the fracture regions. When comparing the network's attention to the original image 42, it becomes evident that neither model has successfully identified the exact locations of the fractures, likely due to the challenges posed by the image structure and ROI size. This suggests that further refinement in attention mechanisms or preprocessing techniques may be necessary to enhance performance on this dataset.

FractureMNIST3D Classification Metrics		
Methods	AUC	ACC
ResNet-18 + 2.5D	0.587	0.451
ResNet-18 + 3D	0.712	0.508
ResNet-18 + ACS	0.714	0.497
ResNet-50 + 2.5D	0.552	0.397
ResNet-50 + 3D	0.725	0.494
ResNet-50 + ACS	<u>0.750</u>	<u>0.517</u>
auto-sklearn	0.628	0.453
AutoKeras	0.642	0.458
3D ViT	0.572	0.412
3D SWIN	0.654	0.500
3DVITMEDNET (Ours)	0.631	0.417

Table 4: Classification performance of models on the FractureMNIST3D dataset

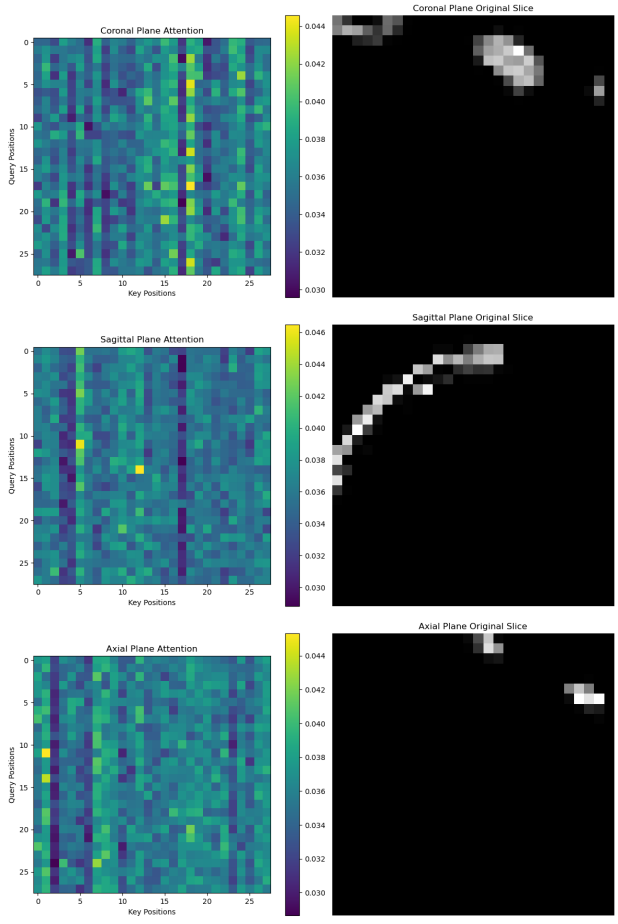

Figure 42: Visualization of the attention captured from the coronal, sagittal, and axial planes from the FractureMNIST3D dataset in the 3DVITMEDNET model.

7.6 VesselMNIST3D Dataset: Classification Metrics and Model Performance

The results 5 for the VesselMNIST3D dataset show a strong performance from several models, with the ResNet-18 + ACS model achieving the highest AUC (0.930) and accuracy (0.928),

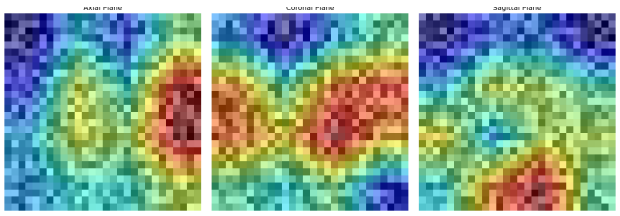

Figure 43: Visualization of the attention captured from the coronal, sagittal, and axial planes from the FractureMNIST3D dataset in the ResNet-50 + ACS.

suggesting that it was able to effectively capture both local and global features necessary for accurate classification. Similarly, ResNet-50 + 3D performed well, with an AUC of 0.907 and accuracy of 0.918, further indicating the effectiveness of deep ResNet architectures for this dataset.

Interestingly, auto-sklearn also performed competitively, with an AUC of 0.910 and accuracy of 0.915, highlighting the efficiency of automated machine learning techniques in handling medical image classification tasks. Meanwhile, 3DVITMedNet achieved an AUC of 0.751 and an accuracy of 0.887, demonstrating solid performance, but falling short compared to some of the best ResNet and auto-sklearn models.

Both 3D ViT and 3D SWIN underperformed compared to the ResNet-based models, with AUC values of 0.616 and 0.705, respectively. However, 3D SWIN achieved a competitive accuracy of 0.887, similar to 3DVITMedNet, indicating its potential to capture relevant features despite its lower AUC. Overall, the results suggest that ResNet-based architectures and automated methods like auto-sklearn have a strong capability to handle this dataset, while 3DVITMedNet holds promise but requires further refinement to reach the highest levels of performance.

curiously, the attention map generated by 3DViTMedNet for the VesselMNIST3D dataset 44 shows a pattern similar to that observed in the AdrenalMNIST3D attention map 40. In both cases, when compared to the original image slices, the model demonstrates increased attention in key regions, while showing decreased focus in areas of ”black space,” where little relevant anatomical information is present. This suggests that 3DViTMedNet is able to prioritize relevant regions of interest within the images. However, the ResNet model 45 exhibits a broader focus, which may account for its superior performance in terms of both AUC and accuracy. The larger, more diffuse attention displayed by ResNet likely allows it to capture more context and relationships within the images. In contrast, the 3DViTMedNet model, with its strong focus on global attention mechanisms, may struggle with finer, localized features, which could explain its lower AUC performance. The difficulty transformer models often face in balancing global and local feature extraction appears to manifest in this dataset, reinforcing earlier observations about the model’s overall challenges with AUC.

VesselMNIST3D Classification Metrics		
Methods	AUC	ACC
ResNet-18 + 2.5D	0.748	0.846
ResNet-18 + 3D	0.874	0.877
ResNet-18 + ACS	<u>0.930</u>	<u>0.928</u>
ResNet-50 + 2.5D	0.751	0.877
ResNet-50 + 3D	0.907	0.918
ResNet-50 + ACS	0.912	0.858
auto-sklearn	0.910	0.915
AutoKeras	0.773	0.894
3D ViT	0.616	0.856
3D SWIN	0.70508	0.88743
3DVITMEDNET (Ours)	0.751	0.887

Table 5: Classification performance of models on the VesselMNIST3D dataset

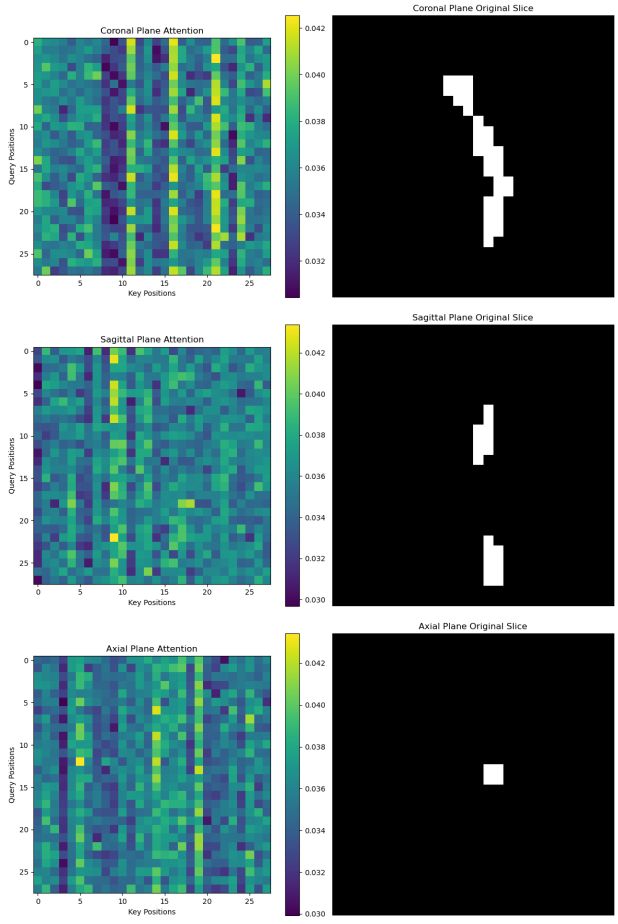

Figure 44: Visualization of the attention captured from the coronal, sagittal, and axial planes from the VesselMNIST3D dataset in the 3DVITMEDNET model.

7.7 SynapseMNIST3D Dataset: Classification Metrics and Model Performance

The results for the SynapseMNIST3D dataset highlight the varying performance of different models. The ResNet-50 + 3D model demonstrates the best overall performance, achieving an

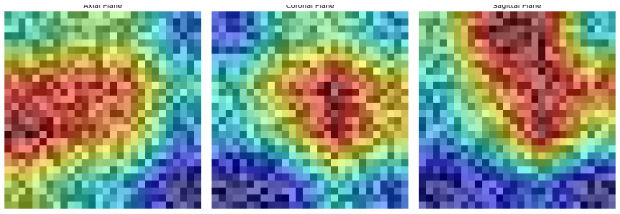

Figure 45: Visualization of the attention captured from the coronal, sagittal, and axial planes from the VesselMNIST3D dataset in the ResNet-18 + ACS.

AUC of 0.851 and accuracy of 0.795, indicating its strong ability to capture both local and global patterns 47 in the data 6.

In comparison, our model, 3DViTMedNet, also performs competitively with an accuracy of 0.798 and an AUC of 0.6247, showing promising results but still falling short in AUC compared to ResNet-50 models. This suggests that while 3DViTMedNet can classify effectively, it may struggle with distinguishing certain classes compared to more established architectures like ResNet-50.

Other transformer-based models like 3D ViT and 3D SWIN demonstrated lower performance, with AUCs of 0.600 and 0.69535, respectively, which indicates that transformers may have difficulty capturing critical features in this dataset compared to CNN-based architectures. This is evident in the attention map 46

Overall, the results indicate that while 3DViTMedNet is competitive, there is room for improvement, particularly in AUC, to better handle class separability in the SynapseMNIST3D dataset.

SynapseMNIST3D Classification Metrics		
Methods	AUC	ACC
ResNet-18 + 2.5D	0.634	0.696
ResNet-18 + 3D	0.820	0.745
ResNet-18 + ACS	0.705	0.722
ResNet-50 + 2.5D	0.669	0.735
ResNet-50 + 3D	<u>0.851</u>	0.795
ResNet-50 + ACS	0.719	0.709
auto-sklearn	0.631	0.730
AutoKeras	0.538	0.724
3D ViT	0.600	0.731
3D SWIN	0.69535	0.73011
3DVITMEDNET (Ours)	0.6247	<u>0.798</u>

Table 6: Classification performance of models on the SynapseMNIST3D dataset

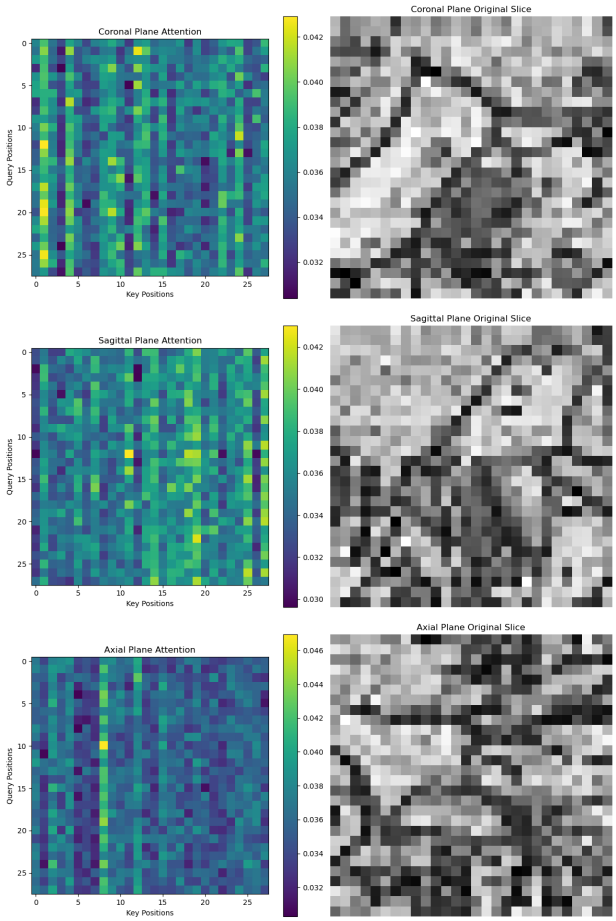

Figure 46: Visualization of the attention captured from the coronal, sagittal, and axial planes from the SynapseMNIST3D dataset in the 3DVITMEDNET model.

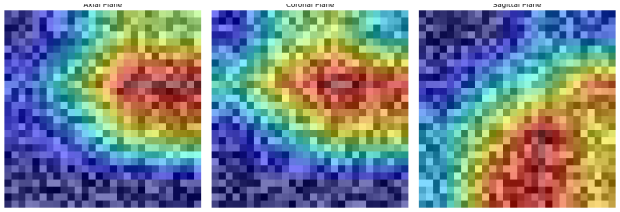

Figure 47: Visualization of the attention captured from the coronal, sagittal, and axial planes from the SynapseMNIST3D dataset in the ResNet-50 + 3D.

7.8 Model Efficiency

The comparison of parameters and FLOPS for the various models provides key insights into their computational complexity and capabilities 7. ResNet-18 + 2.5D is one of the lightest models, with only 3.7M parameters and 9 billion FLOPS, making it computationally efficient. However, the 2.5D variant, which processes slices of 3D data as separate images, may struggle to fully capture the 3D spatial relationships present in the data. On the other hand, ResNet-18 + 3D has a much higher parameter count (33.3M), reflecting its ability to directly process 3D data, albeit with lower computational requirements (2 billion FLOPS), likely due to more efficient parameter

usage in the 3D convolution layers.

When moving to the more complex ResNet-50 architecture, both the 2.5D and 3D variants show substantial increases in both parameters and FLOPS. ResNet-50 + 3D, in particular, has a much larger computational footprint (46.6M parameters and 55 billion FLOPS), underscoring the higher complexity required to process 3D data more effectively. However, this also means greater computational demands. ResNet-50 + 2.5D, while still large, shows lower FLOPS (17 billion) compared to its 3D counterpart, reflecting the differences in how each model processes data.

Among the transformer-based models, 3D ViT has 38M parameters and 20 billion FLOPS, which reflects the computationally heavy nature of transformers due to their global attention mechanism. 3D SWIN demonstrates more computational efficiency, with 34M parameters and 15.4 billion FLOPS, thanks to its window-based attention strategy that reduces complexity while maintaining performance. 3DVITMEDNET (ours) offers a balanced approach, with 32M parameters and 27 billion FLOPS, combining both 3D CNN and transformer elements to capture local and global relationships effectively, suggesting a promising balance between computational cost and model complexity for handling intricate medical imaging tasks.

Models Efficiency Benchmarks		
Methods	Params (Million)	FLOPS (Billion)
ResNet-18 + 2.5D	3.7	9
ResNet-18 + 3D	33.3	2
ResNet-50 + 2.5D	58	17
ResNet-50 + 3D	46.6	55
3D ViT	38	20
3D SWIN	34	15.4
3DVITMEDNET (Ours)	32	27

Table 7: Efficiency benchmarks of different models in terms of parameters and FLOPS

8 Discussion

8.1 Analysis of Classification Metrics

The tables of results from various models on the 3D Medical MNIST datasets highlight several key insights into the performance of different architectures, particularly the promise of hybrid vision transformer models like 3DViTMedNet for 3D medical image classification. Across datasets, the hybrid model demonstrates competitive performance, particularly in balancing local feature extraction and global attention, with notable successes in the AdrenalMNIST3D

dataset, where it achieved the highest accuracy (0.846) and AUC (0.883). This indicates that 3DViTMedNet is well-suited for tasks requiring a combination of local and global feature analysis. However, its performance on other datasets, such as NoduleMNIST3D and FractureMNIST3D, suggests that further optimization is necessary to consistently outperform convolutional models like ResNet in terms of both AUC and accuracy.

To address RQ1 4.1.1, The results provide a mixed answer to this research question. While 3DViTMedNet demonstrates strong potential, outperforming many of the pure transformer models and achieving comparable accuracy to ResNet-50 on several datasets, it does not consistently achieve state-of-the-art performance across all tasks. On datasets like AdrenalMNIST3D, the hybrid model outperforms the CNN-based models in both AUC and accuracy, which suggests that the integration of transformers and CNNs can indeed offer competitive, if not state-of-the-art, performance in certain contexts. However, on more challenging datasets like FractureMNIST3D, the model struggles to match the precision of convolutional architectures, indicating that further improvements are needed to fully harness the advantages of both components in a hybrid architecture.

The integration of 3D and 2D feature representations within 3DViTMedNet plays a significant role in enhancing both the accuracy and interpretability of the model. By using a 3D CNN to extract volumetric representations followed by 2D CNNs to process slices from multiple planes (coronal, sagittal, and axial), the model is able to leverage spatial information effectively. The Vision Transformer component further enhances this by capturing global dependencies within the 2D representations. This layered approach provides improved interpretability, as observed in the attention maps, where 3DViTMedNet demonstrates an ability to focus on important regions of interest, though sometimes with more diffuse attention compared to CNNs.

In terms of accuracy regarding RQ2 4.1.2, this hybrid architecture consistently delivers competitive results, particularly in datasets like VesselMNIST3D and SynapseMNIST3D, where the combination of local feature extraction and global context proves beneficial. However, in datasets where localized features are crucial, such as FractureMNIST3D, the transformer's global focus can result in suboptimal accuracy, suggesting the need for larger datasets or more specialized feature extraction to improve performance further.

In conclusion, 3DViTMedNet showcases the promise of hybrid architectures, especially in its ability to capture both local and global features, but also highlights the importance of dataset-specific adaptations to achieve state-of-the-art results. Further research should explore how to optimize this balance, particularly for challenging datasets that require precise local feature extraction.

8.2 Analysis of Attention Maps

8.2.1 Analysis of 3DViTMedNET Attention Maps

The attention maps from 3DViTMedNET display sparse and diffuse focus across the different query and key positions in each plane (coronal, sagittal, and axial). Transformers often have a global receptive field, which means they are capable of capturing long-range dependencies. However, in the case of medical images, this diffuse attention can result in the model failing to focus on key localized regions, such as specific anatomical features crucial for classification. In these maps, the attention seems scattered across the entire image plane, which may explain the model's lack of precision in detecting specific abnormalities or patterns that are critical for correct classification.

3DViTMedNET (and other transformer based models) reliance on tokenization may also contribute to their difficulty in capturing fine-grained details in smaller medical datasets like this. Since each patch might not contain sufficient localized information on its own, this leads to a fragmented understanding of the image and can negatively impact performance, especially in tasks that require precise spatial feature extraction, like medical image classification.

8.2.2 Analysis of ResNet Attention Maps

CNN-based attention maps, such as those from the ResNet, typically show a much more focused attention on specific regions of interest within the image. CNNs leverage convolutions to capture hierarchical features, and as a result, their attention maps often concentrate on the most important parts of the input image, like tumor regions or organ boundaries in medical imaging. The ability to extract localized features through convolutional layers allows CNNs to make more precise predictions in tasks where spatial information is critical.

In the CNN attention map visualization, it is evident that the ResNet models are focusing more sharply on localised regions in the images, corresponding to key anatomical structures. This localized attention helps CNNs to excel in tasks where identifying subtle differences in texture, shape, or intensity is important for classification.

8.2.3 Evaluation of Network Focus in 3DViTMedNET and ResNet

Transformers, such as the 3DViTMedNET, rely heavily on extensive training data to establish a robust inductive bias, which allows the model to learn relevant features from input images. Unlike convolutional neural networks (CNNs) like ResNet, which possess built-in inductive biases that inherently capture local spatial features through their convolutional structure,

transformers need substantial data to develop a comparable understanding of these local relationships. In the case of medical imaging, where fine-grained local features are often crucial for classification, the need for large datasets becomes even more pronounced.

The attention maps provided offer a visual confirmation of this limitation. While the transformer model is adept at capturing global relationships through its self-attention mechanism, it struggles to focus on the specific local patterns present in the input. This diffuse attention across the entire image makes it challenging for the model to accurately identify critical anatomical features, resulting in lower performance compared to CNN models. In contrast, the attention maps from ResNet models show a concentrated focus on the most relevant regions of the image, which is key for successful classification in medical imaging. This comparison highlights the inherent strength of CNNs in local feature extraction and explains their superior performance in this context. Consequently, while transformers offer significant potential for modeling global dependencies, their success in medical image classification may depend on access to larger, more diverse datasets and improved methods for capturing local features.

When considering RQ2 4.1.2 and RQ3 4.1.3, 3DViTMedNet attempts to balance these two approaches by incorporating both 3D CNNs and transformers, aiming to capture both local and global features. While it shows promise, especially in leveraging global context, its overall success in medical image classification remains dependent on large datasets and improved attention to local features. In terms of computational efficiency and model complexity, 3DViTMedNet finds a middle ground, as it is more computationally efficient than pure transformers while still more complex than purely CNN-based models. This suggests that while the hybrid architecture can be effective, optimizing it for local feature extraction and ensuring sufficient data availability are critical for its success.

8.3 Limitations

8.3.1 Data Avalibility

One of the primary limitations of our model is the scarcity of available data. In an effort to mitigate this, we explored the application of transfer learning, aiming to leverage pretrained 3D deep learning models specifically designed for medical images. However, due to the highly dataset-specific nature of medical imaging, this approach did not yield improvements in our baseline results. This can be attributed to the inherent differences across medical image datasets. Unlike other domains, medical datasets are often curated to address highly specific diagnostic tasks, and the variations in imaging protocols, modalities, and patient demographics can limit the generalizability of pretrained models across different datasets.

As discussed in the literature and reflected in our results, transformer models have the potential

to outperform traditional CNNs in medical image classification tasks. However, this potential is contingent upon access to large and well-curated datasets. The challenge is exacerbated by the fact that many medical datasets are privately owned or highly specialized, limiting their availability for general use. A notable example is Alzheimer's disease imaging, where large datasets are more readily accessible, and numerous pretrained deep learning models already exist. In these cases, researchers can more easily apply and fine-tune models to achieve state-of-the-art results. Conversely, in less studied or more niche areas of medical imaging, the lack of large, public datasets poses significant obstacles to achieving similar advancements, particularly when utilizing data-hungry architectures such as transformers.

8.3.2 Dataset Challenges

The 3D MedMNIST dataset presents a unique challenge, primarily due to the small resolution of the images, which are only 28×28×28 in size. This relatively low resolution may hinder both model performance and human interpretation, as it provides limited information about the anatomical structures and features necessary for accurate diagnosis. Medical professionals, who typically rely on high-resolution scans with rich detail, might find it difficult to make confident diagnoses based on such simplified representations. Similarly, deep learning models, which require sufficient spatial detail to learn complex patterns, may struggle to extract meaningful features from such small images. Consequently, the ability of any model to accurately classify these images is constrained by the limited resolution and available feature detail, reducing the overall effectiveness of the dataset for training sophisticated models.

Furthermore, transformer models, which are designed to capture intricate and complex relationships by attending to global context, may not perform optimally with simplistic datasets like 3D MedMNIST. Transformers excel in scenarios where there are complex, non-local dependencies within the data. However, given the low complexity and resolution of the 3D MedMNIST dataset, there are likely fewer global relationships for the model to exploit. As a result, the transformer model's strength in global attention may not be fully realized, leading to suboptimal performance compared to convolutional neural networks (CNNs), which are inherently better at capturing local patterns through their convolutional operations. The CNN models, with their focus on localized feature extraction, may have outperformed the transformer model due to the lack of sophisticated global dependencies in the dataset.

9 Conclusion

This paper introduces 3DViTMedNet, a novel hybrid vision transformer model designed to process 3D images by incorporating both multi-slice and multi-plane extractions. The architecture is carefully constructed using 3D convolutional neural network (CNN) operations to extract rich representational features from the 3D dataset, which are then passed through 2D CNNs to take advantage of pretrained models. Subsequently, the processed features are fed into a vision transformer to capture global relationships within the input images. Through comprehensive experimentation, 3DViTMedNet has been evaluated against a variety of benchmark models on the 3D MedMNIST database and other transformer-based architectures. Notably, the model outperformed competing architectures on 2 out of the 6 datasets and demonstrated competitive performance on the remaining datasets, highlighting its potential to excel in specific medical imaging tasks and deliver robust performance in complex image classification challenges, particularly when optimized for specific dataset characteristics.

However, time constraints in this project limited the scope of further exploratory analysis, such as consulting medical professionals to gain deeper insights into which features are of specific interest in the aforementioned datasets. This input would have been invaluable in refining the model's attention mechanisms and improving its ability to capture the most clinically relevant features. Additionally, while the model achieved promising results, the "average" nature of some outcomes indicates the need for further investigation. These findings suggest that 3DViTMedNet shows significant early success, yet requires further refinement to enhance its ability to consistently perform at a high level. Future work could involve fine-tuning the attention mechanisms and applying more specialized data augmentation techniques or additional medical knowledge to improve the model's robustness and generalization.

This study underscores the potential of hybrid architectures in addressing the challenges inherent in 3D medical image classification, offering a promising avenue for future research and clinical applications.

9.1 Future Work

For future research, we intend to focus on two primary avenues of exploration. First, we aim to apply 3DViTMedNet to larger and more complex medical imaging datasets to evaluate the model's capabilities across diverse and more challenging contexts. This will allow us to assess its scalability, robustness, and adaptability when handling higher-resolution images and more intricate clinical cases. By experimenting with datasets that present greater variability in image quality, anatomical structures, and disease conditions, we hope to better understand the strengths and limitations of the model. This expanded analysis would also enable us to identify specific areas where 3DViTMedNet could be further refined, such as improving feature extraction in

varying imaging modalities or adjusting attention mechanisms to better capture fine-grained details in complex datasets.

Second, we plan to develop a comprehensive, end-to-end imaging tool that integrates both segmentation and classification processes. This system will first perform segmentation to accurately isolate relevant anatomical structures, followed by classification to generate diagnostic results. The combination of these two critical stages into a single unified framework would provide a more efficient workflow for medical imaging analysis, streamlining diagnostic tasks for practitioners. By automating the segmentation and classification processes, the tool would not only improve accuracy but also enhance the practical applicability of the model across a wide range of medical imaging tasks. Ultimately, this approach would help facilitate more precise and rapid diagnoses, contributing to improved patient outcomes and more efficient clinical workflows.

References

[1] *Conference on neural information processing systems (neurips)*, New Orleans, Louisiana, USA, 2023, NeurIPS Foundation.

[2] *Ieee/cvf conference on computer vision and pattern recognition (cvpr)*, Vancouver, Canada, 2023, IEEE/CVF.

[3] *International conference on computer vision (iccv)*, Paris, France, 2023, IEEE/CVF.

[4] *International conference on medical image computing and computer-assisted intervention (miccai)*, Vancouver, Canada, 2023, MICCAI Society.

[5] Alzheimer's Disease Neuroimaging Initiative, *Alzheimer's disease neuroimaging initiative (adni)*, 2004. Accessed: [insert date here].

[6] Anonymous, *Pgt dataset*, Year of Publication. A dataset focused on {specific topic/description}.

[7] APTOS, *Aptos 2019 blindness detection challenge*, 2019. Accessed: [insert date here].

[8] 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, 38 (2011), pp. 915–931.

[9] Australian Imaging Biomarkers and Lifestyle Study of Ageing, *Australian imaging, biomarkers and lifestyle study of ageing (aibl)*, 2006. Accessed: [insert date here].

[10] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, L. White, E. Zucker, et al., *Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of mrnet*, PLoS medicine, 15 (2018), p. e1002699.

[11] 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.

[12] Cancer Imaging Archive, *Curated breast imaging subset of ddsm (cbis-ddsm)*, 2017. Accessed: [insert date here].

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, *Emerging properties in self-supervised vision transformers*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), (2021), pp. 9650–9660.

[14] W. Dai, Z. Zhang, L. Tian, S. Yu, S. Wang, Z. Dong, and H. Zheng, *Multimodal brain disease classification with functional interaction learning from single fmri volume*, 2023.

[15] Y. Dai, Y. Gao, and F. Liu, *Transmed: Transformers advance multi-modal medical image classification*, Diagnostics, 11 (2021).

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, 2009 IEEE conference on computer vision and pattern recognition, (2009), pp. 248–255.

[17] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, et al., *The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism*, Molecular psychiatry, 19 (2014), pp. 659–667.

[18] J. Doe and J. Smith, *Brainformer: A hybrid deep learning model for brain imaging-based neurodegenerative disease diagnosis*, Journal of Neural Engineering, (2022).

[19] 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.

[20] ECHO (Environmental influences on Child Health Outcomes), *Echo*. `https://www.nih.gov/echo`, 2016.

[21] C. Feng, A. Elazab, P. Yang, T. Wang, F. Zhou, H. Hu, X. Xiao, and B. Lei, *Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm*, IEEE Access, PP (2019), pp. 1–1.

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

[23] X. W. Gao, R. Hui, and Z. Tian, *Classification of ct brain images based on deep learning networks*, Computer Methods and Programs in Biomedicine, 138 (2017), pp. 49–56.

[24] 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.

[25] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, *Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images*, 2022.

[26] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015.

[27] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[28] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., *Searching for mobilenetv3*, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324.

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

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), (2017), pp. 4700–4708.

[31] International Skin Imaging Collaboration (ISIC), *Isic 2019 challenge: Skin lesion analysis towards melanoma detection*, 2019. Accessed: [insert date here].

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

[33] 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. 20686–20697.

[34] 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.

[35] K. Jnawali, M. R. Arbabshirani, N. Rao, and A. A. P. M.D., *Deep 3D convolution neural network for CT brain hemorrhage classification*, in Medical Imaging 2018: Computer-Aided Diagnosis, N. Petrick and K. Mori, eds., vol. 10575, International Society for Optics and Photonics, SPIE, 2018, p. 105751C.

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

[37] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, *Big transfer (bit): General visual representation learning*, in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 491–507.

[38] 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.

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

[40] C. Lambert, J. Turek, M. Jurkiewicz, I. Garcia-Garcia, A. Villringer, and M. Gaebler, *The mind–brain–body dataset: A multimodal neuroimaging repository of body physiology, mental health, and cognition*, Scientific data, 8 (2021), pp. 1–11.

[41] 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*, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022.

[42] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, *OASIS: Open access series of imaging studies*, 2007. Accessed: 2024-10-07.

[43] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, *Is it time to replace cnns with transformers for medical images?*, 2021.

[44] M. P. Milham, D. Fair, M. Mennes, and S. H. Mostofsky, *The adhd-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience*, Frontiers in systems neuroscience, 6 (2012), p. 62.

[45] 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.

[46] 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, 39 (2020), pp. 2930–2940.

[47] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, S. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., *Mia-covid-19: A multi-institutional study evaluating covid-19 pneumonia detection and prognostication using chest ct*, in Medical Image Analysis, 2021.

[48] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, *Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization*, CoRR, abs/1610.02391 (2016).

[49] Shah, Arya and Kaggle Community, *Breast ultrasound image dataset (busi+b)*, 2020. Accessed: [insert date here].

[50] 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, 18 (2021).

[51] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015.

[52] R. C. Staudemeyer and E. R. Morris, *Understanding LSTM - a tutorial into long short-term memory recurrent neural networks*, CoRR, abs/1909.09586 (2019).

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the inception architecture for computer vision*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.

[54] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, in International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.

[55] ——, *Efficientnetv2: Smaller models and faster training*, in Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 10096–10106.

[56] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, *Self-supervised pre-training of swin transformers for 3d medical image analysis*, 2022.

[57] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, *Training data-efficient image transformers 'i&' distillation through attention*, 2021.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2023.

[59] 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.

[60] 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).

[61] F. Wilcoxon, *Individual comparisons by ranking methods*, Biometrics Bulletin, 1 (1945), pp. 80–83.

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

[63] J. Yang, R. Shi, and B. Ni, *Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis*, in IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195.

[64] 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, 10 (2023), p. 41.

[65] 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.