Rethinking Biomedical Image Classification Through Vision Transformers and Foundation Models on MedMNISTv2

Roza Dler Abdalla

Master of Science in Artificial Intelligence

from the

University of Surrey

School of Computer Science and Electrical and Electronic Engineering

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

September 2024

Supervised by: Prof. Gustavo Carneiro

DECLARATION OF ORIGINALITY

I confirm that the project dissertation I am submitting is entirely my own work and that any material used from other sources has been clearly identified and properly acknowledged and referenced. In submitting this final version of my report to the JISC anti-plagiarism software resource, I confirm that my work does not contravene the university regulations on plagiarism as described in the Student Handbook. In so doing I also acknowledge that I may be held to account for any particular instances of uncited work detected by the JISC anti-plagiarism software, or as may be found by the project examiner or project organiser. I also understand that if an allegation of plagiarism is upheld via an Academic Misconduct Hearing, then I may forfeit any credit for this module or a more severe penalty may be agreed.

MSc Dissertation Title: Rethinking Biomedical Image Classification Through Vision Transformers and Foundation Models on MedMNISTv2

Author Name: Roza Dler Abdalla

Author Signature: Roza Dler Abdalla   Date: Sep 17, 2024

Supervisor's name: Prof. Gustavo Carneiro

WORD COUNT

Number of Pages: 91

Number of Words: 21988

ABSTRACT

Medical image classification is a critical task in healthcare, aiding in disease diagnosis and treatment planning. This project rethinks the benchmarks established by MedMNIST by exploring the performance of newer deep learning models, including Vision Transformers (ViTs), hybrid CNN-transformer models like MedViT, and vision-language foundation models such as MedCLIP. Experiments were conducted using three diverse datasets from the MedMNIST v2 collection: PneumoniaMNIST (chest X-rays), PathMNIST (colon pathology slides), and VesselMNIST3D (brain MRA images). The ViT experiments revealed that fine-tuning with larger batch sizes and image sizes allowed the model to surpass the existing benchmarks, particularly for PathMNIST, where the global attention mechanism of ViTs captured intricate spatial relationships in the images. However, on the highly imbalanced and small 3D dataset, VesselMNIST3D, ViT struggled to generalize, whereas the joint architecture of ViT and CNN performed similarly to the benchmark, indicating the need for further investigation into handling complex 3D medical data. MedViT, designed as a hybrid model combining CNNs' local feature extraction with the global attention of transformers, showed promise by performing well on simpler tasks like PneumoniaMNIST. It worked effectively with simpler hyperparameters, reducing the need for extensive tuning and making it an accessible option for other medical imaging applications. The exploration of foundation models showed that fine-tuning MedCLIP achieved the highest results for the larger, multiclass PathMNIST, showcasing the power of multimodal pretraining. Using MedCLIP's vision-only model highlighted its effectiveness even without the inclusion of language inputs, suggesting the potential of integrating language and its impact on the domain. Additionally, zero-shot prompt-based classification was explored using MedCLIP, demonstrating that when pre-trained on medical text, the model could classify efficiently with medical-context prompts, performing exceptionally well on PneumoniaMNIST. However, it struggled with PathMNIST, due to a lack of pretraining on pathology-specific data. This dissertation provides a comprehensive evaluation of these newer models, emphasizing the significance of model selection, pretraining, and prompt engineering in advancing AI-assisted biomedical imaging and guiding future work in the field.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Biomedical image classification is an area within computer vision that focuses on analyzing and interpreting medical images, such as X-rays, ultrasound, MRIs, and histopathology slides, to assist in the process of diagnostics, exploratory clinical tasks, academic research, and treatment. This field is essential in healthcare, where accurate and efficient image analysis can significantly impact patient care. The use of machine learning, specifically deep learning has enabled significant advances in this domain, allowing for more reliable, automated classification of complex medical images, while providing aid for medical professionals, now needed more than ever due to the rising computed-aided testings that result in a larger volume of biomedical imaging that gets produced and need to be interpreted and analyzed by healthcare professionals [2]. The purpose of this project is to explore and extend the benchmarks in biomedical image classification using the MedMNIST dataset collection. The work involves experimental research, where various deep learning models such as vision transformers and vision-language foundation models, are implemented, fine-tuned, and evaluated on these standardized datasets. Through this experimental approach, the project aims to identify new pathways for achieving robust and generalizable performance in biomedical image analysis.

1.1 Background and Context

The field of medical imaging has witnessed significant advancements in recent years, through the integration of Machine Learning (ML) and Deep Learning (DL)techniques. These methods have demonstrated performance comparable to medical experts in specific tasks and are beginning to receive clinical certifications [29]. The importance of biomedical image analysis, paired with the magnitutude of newer and ever-growing methods and techniques in medical imaging, has inspired the need for research in Artificial Intelligence (AI) assisted biomedical imaging techniques and algorithms. The use of deep learning has brought advancements to medical image classification, a subset of medical image analysis. As it enables the detection and classification of complex patterns within medical images that might otherwise require extensive time and expertise to interpret [33]. However, despite the growing success of deep learning in this field, much of the research has been concentrated on a limited set of popular datasets, resulting in an overemphasis on achieving state-of-the-art (SOTA) performance on the same limited, influential datasets [9]. This focus has led to two major challenges: an oversaturation of results on the same datasets, limiting their generalizability, and the development of increasingly complex models designed solely to surpass existing benchmarks, which may not translate to real-world medical applications [9]. In response, recent works have focused on creating diverse and standardized benchmark datasets, such as the MedMNIST dataset collection to facilitate the development and evaluation of new DL techniques and models, prioritizing generalizability and research towards new benchmarking spanning across different datasets, image sizes, tasks and modalities [51]. MedMNIST aids in addressing these issues by introducing a collection of lightweight, standardized biomedical image datasets designed for classification tasks. By offering a diverse range of 2D and 3D datasets in a compact MNIST-like format, MedMNIST provides researchers with an accessible sset of uniform and standardized dataset collection to experiment with new deep learning models without requiring domain-specific knowledge [25], [51].

Building upon these efforts, this project aims to work towards extending the benchmarking work initiated by MedMNIST by further diversifying the benchmark landscape and evaluating the potential of different and newer models, such as vision transformers and vision-language foundation models. This project focuses on three distinct datasets from the MedMNIST v2 collection: PathMNIST, derived from histopathology slides and representing the largest dataset in the collection; PneumoniaMNIST, consisting of chest X-ray images; and VesselMNIST3D, which involves 3D brain MRA images. This work aims to provide a comprehensive evaluation of the selected models and techniques, showcasing their potential for generalization across various medical imaging domains, as well as showcasing potentiality of alternative approaches such as zero-shot prompt-based classifications utalising large vision-language foundation models.

1.2 Objectives

The primary objective of this project is to work towards extending and diversifying the benchmarks of the MedMNIST v2 dataset collection by exploring the potential of different models for biomedical image classification. The project aims to address the limitations of existing benchmarks by experimenting with state-of-the-art models, including vision transformers (ViTs) and vision-language foundation models pre-trained on medical data. The specific objectives are:

- To verify the current state-of-the-art (SOTA) benchmarks on three selected MedMNISTv2 datasets: PneumoniaMNIST, PathMNIST, and VesselMNIST3D by reproducing the baseline models (ResNet-18, ResNet-50, and ResNet-18 + ACS convolution for 3D).
- To propose and fine-tune vision transformer (ViT) models for 2D and 3D biomedical image classification tasks and evaluate its performance on the selected datasets.
- To investigate the effectiveness of pre-trained models on biomedical data by fine-tuning (MedViT) on the three datasets and assessing the improvements in classification performance.
- To explore the performance of vision-language foundation models (MedCLIP) pre-trained on medical data, including their application in fine-tuning and zero-shot prompt-based classification.
- To evaluate and compare the impact of different model architectures (ResNet, ViT, and MedCLIP) on generalization and classification accuracy across diverse data modalities in MedMNIST v2, providing insights into model selection for future biomedical image analysis research.

1.2.1 Summary of Tasks

To fulfill these objectives, the following tasks were undertaken:

- Dataset Selection and Preparation: Selected three datasets from MedMNIST v2 (PneumoniaMNIST, PathMNIST, and VesselMNIST3D) and ensured alignment with MedMNIST's preprocessing standards.
- Benchmark Reproduction: Implemented the current SOTA benchmark models (ResNet-18, ResNet-50, and ResNet-18 + ACS convolution for 3D) for the selected datasets and verified their performance (accuracy and AUC) against the reported benchmarks.
- Model Implementation: Proposed and fine-tuned a vision transformer (ViT_B/16) model for the 2D datasets and a custom 3D ViT model for VesselMNIST3D.
- Pre-trained Model Fine-tuning: Fine-tuned MedViT, a vision transformer pre-trained on medical data, on the selected 2D datasets to explore the impact of transfer learning on a ViT pretrained on medical data.

- Foundation Model Experiments: Fine-tuned MedCLIP, a vision-language model pre-trained on medical data, and conducted zero-shot prompt-based classification to evaluate its generalization capabilities.
- Results Analysis: Analyzed the performance of each model to identify strengths, weaknesses, and areas of potential improvement, contributing new insights into the application of these models for biomedical image classification.

1.3 Achievements

This project successfully achieved its primary objectives, providing valuable insights into the application of newer deep learning models for biomedical image classification using the MedMNIST v2 datasets. Firstly, the SOTA benchmarks were effectively reproduced across PneumoniaMNIST, PathMNIST, and VesselMNIST3D, establishing a solid foundation for evaluating other models. The fine-tuning of Vision Transformers (ViTs) resulted in surpassing traditional benchmarks, particularly for PathMNIST, where larger batch sizes and image dimensions enhanced performance. MedViT, a hybrid CNN-transformer model, demonstrated strong results, especially on PneumoniaMNIST, showing that simpler hyperparameters can still achieve robust performance, highlighting its adaptability for medical imaging tasks.

Additionally, MedCLIP was fine-tuned to achieve outstanding results on the larger PathMNIST dataset, confirming the potential of vision-language models in complex medical image classification. The exploration of zero-shot prompt-based classification using MedCLIP provided further insights, revealing its strengths in scenarios with relevant medical prompts, particularly for PneumoniaMNIST. While there are areas for improvement, such as refining the approach for small, imbalanced 3D datasets and further utilizing MedCLIP's full multimodal capabilities, this project contributes significantly to understanding how diverse models can be applied and extended in the biomedical imaging domain.

1.4 Overview of Dissertation

This dissertation begins with an Introduction, establishing the context of biomedical image classification and its growing importance in healthcare, especially with the rise of deep learning (DL) methodologies. It presents the aims of this research, focusing on expanding MedMNIST's benchmarking by exploring newer models, including ViTs and vision-language foundation models, as well as examining alternative approaches like zero-shot classification.

Following the introduction, the Background and Context chapter provides an in-depth overview of the domain, detailing the challenges in medical imaging and the increasing need for AI-assisted solutions. It introduces the concepts of deep learning within medical imaging, covering key methodologies such as Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and vision-language models. Relevant literature and related studies are discussed to highlight the current landscape of medical image classification and situate this project within ongoing research, ending with the most directly related work to this study.

The Methodology section outlines the datasets chosen for this study, describes preprocessing steps, and details the architectures of the models used. It covers the training setups and evaluation metrics employed, ensuring consistency with MedMNIST's standards to facilitate fair comparison. This section provides a comprehensive overview of the experimental setup, preparing the reader for the results and analyses that follow.

The dissertation then moves to Experimental Results and Analysis, where each dataset is systematically explored with each model. The outcomes of the experiments are presented, followed by an analysis of each model's performance across the datasets. Here, key findings are discussed, such as the performance of ViTs, the strengths of MedViT in hybrid learning, and MedCLIP's versatility in fine-tuned and zero-shot scenarios. The project concludes with a Conclusion and Future Work section, summarizing the contributions made and outlining future research directions, including more extensive zero-shot classification experiments, robust handling of 3D data, and further exploration of foundation models in medical imaging. This structure provides a clear narrative, guiding the reader through the research journey from foundational concepts to advanced model evaluations.

2 BACKGROUND THEORY AND LITERATURE REVIEW

2.1 Medical Imaging and AI

Medical Imaging is a strong pillar of healthcare, providing medical professionals with visual representations of the human body for diagnostic, treatment, and exploratory purposes [52]. various techniques and devices are used for medical imaging such as X-rays, computed tomography (CT), Magnetic resonance imaging (MRI), ultrasound, and many more [17]. Such different technologies and techniques in capturing parts of the human body mean there are multitudes of data modalities involved in medical imaging. For example, X-ray imaging produces 2D grey-scale images of structures based on tissue density, commonly used for examining bones, chest, and dental structures [17]. Whereas imaging techniques such as magnetic resonance angiography (MRA), based on Magnetic Resonance Imaging (MRI), focuses on imaging blood vessels in the brain, which could be used to detect aneurysms, and other vascular abnormalities, as well as for monitoring conditions like stroke and atherosclerosis [17], [5]. Brain MRA then produces detailed 3D images of blood vessels in the brain, showcasing the vascular structures rather than general tissue anatomy. Colon Pathology imaging is another example, showing the variety and complexity of different medical imaging modalities. As Colon pathology involves multiple modalities, each with its own unique characteristics and data types. For example Histopathology, a crucial modality for diagnosing colon pathology, specifically cancer, involves examining tissue samples under a microscope. This produces cellular-level high-resolution 2D images, that are often coloured due to stains from Hematoxylin and Eosin [30]. They provide digitized histopathology slides, showing cellular-level details often not visible in other imaging types, and their interpretation relies heavily on colour and texture patterns, requiring high level of expertise and careful consideration from medical professionals [18].

The diversity and complexity of medical imaging, coupled with their crucial role in diagnostics, treatment, and research, makes their interpretation a challenging and delicate task, often undertaken by medical professionals utilising their skills, expertise, and medical references [43]. In the last decades, due to advancements in non-invasive ways for diagnostics and treatment, the volume of medical imaging data produced is on the rise [2]. This has increased the pressure on medical professionals to analyse, interpret, and process a rising number of medical imaging across different modalities and for a variety of purposes. This, in turn, results in human expert fatigue, and often slows down patient care. As much of physicians' time would be spent analysing the different imaging and test results obtained from patients, thereby reducing time spent with patients [4]. Therefore, there was a need for innovation with the goal of aiding medical professionals with medical imaging analysis.

With the advancement of Artificial Intelligence, its potential to revolutionize medical imaging was recognized quickly [43]. As AI's capabilities in pattern recognition, processing, analysis, and classification were recognized, so was its application in the medical imaging domain. Specifically, a crucial part of medical image analysis that AI has revolutionized is classifications. Medical image classification involves categorizing medical imaging data into specific categories, often used for the purpose of diagnosing diseases, abnormalities, or different types, sections, and parts of a condition. For example, images of chest X-rays could be classified to detect the presence or absence of a certain disease like Pneumonia [19].

Traditional approaches in image classification often relied on manual feature extraction, such features were defined by domain experts. Features were selected based on edge, shape, texture, and colour. This created a narrow vision of what features of a certain task should look like, making such early models not generalize and adapt to different tasks and modalities [19]. However, Deep learning has revolutionized this process, by allowing features to learn through the process of 'training' models, to analyse a large number of images and find patterns to create its own unique features that are ever-changing [43]. By automating medical image classification, DL algorithms can aid healthcare providers by streamlining their workflow, existing as an 'extra set of eyes', which serves a role in the credibility verification of reported data [52]. In the coming sections, a number of these deep learning methodologies and models will be explored, and their applications in medical image classification will be highlighted.

2.2 Convolutional Neural Networks In Medical Image Classification

2.2.0.1 Introduction to Deep Learning

Deep learning refers to a subset of machine learning where multi-layered artificial neural networks are used to find patterns in data automatically with a process known as 'Representation Learning'. Traditional machine learning methods use hand-crafted feature extraction, this task requires human expertise and careful consideration to ensure the machine learning method can extract patterns and features from specific data, Examples of features that could be crafted for the methods to extract are edge, shape, texture, line, and colour [11], [26]. On the other hand, deep learning does not require hand-crafted feature extraction methods, as they learn data representations through processing layers. This idea of representation learning has revolutionized the applications of AI, because it allows deep learning models to process and handle vast amounts of data from a multitude of modalities, without needing domain-specific or task-specific feature extraction [26]. This means that deep neural networks can take raw input data, for example, an image, and learn representations from this image through multiple processing layers, each layer 'extracting' increasingly more abstract features. The abstraction refers to the hierarchical manner where the lower, earlier layers can detect low-level features such as edges, lines, and shapes, and as the layers progress, so does the level of abstraction. Where the higher layers are able to recognize more abstract and complex representations such as silhouettes [11]. This hierarchical learning from data without the need for manual feature engineering is the core reason for the successful applications of deep learning in a multitude of domains. Further, another notable breakthrough of deep learning is the ability to scale up with larger datasets and computational powers, techniques such as backpropagation, allow the models to update their weights based on error rates during training, these weight updates are what lead to the ultimate 'learning' of the models during training. These attributes and advantages make deep learning a suitable candidate for medical image analysis, where often subtle and fine-grained pattern recognition is essential for analysis [26].

2.2.0.2 Neural Networks

Inspired by the neurons in the human brain, neural networks refer to the interconnected network of 'nodes', sometimes referred to as neurons, organized in layers. The layers consist of an input layer, what could be perceived as the start of the neural network, where input data is fed to the network in this layer, then one or more internal layers, called 'hidden layers', these layers are responsible for the processing of the input data. Finally, the last layer is the output layer, responsible for producing the final result of the neural network's task [26].

In neural networks, each node is interconnected with preceding nodes. These connections between nodes have associated weights, these weights are the variables that get updated as the model learns representation throughout the training process. The weight updates are done through a process called Backpropagation, introduced by Rumelhart, Hinton, and Williams in 1986 [39]. Backpropagation computes the gradients of the error for the weights, this means the backpropagation algorithm determines how much each weight associated with each connection should be updated to reduce the error rate between the predicted outcome and the real outcome.

Neural network's ability to learn hierarchical representations of data, as information flows through the layers, each layer increasingly learns more abstraction from the input data. Backpropagation facilitates this learning to ensure the network iteratively updates its weights to have the highest confidence, in other words, the smallest error rate between the predicted outcome and the expected outcome [11]. This hierarchical learning is particularly useful for complex tasks such as image classifications, setting the stage for more advanced research and architectures such as Convolutional Neural Networks (CNNs).

2.2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a specialized type of neural network that is able to process data in ways that maximize the representation learning operated by neural networks. CNN's architecture and components makes it effective for image-based downstream tasks, specifically, the two key principles that make CNN effective for inputs such as images are what is referred to as sparse connections and weight sharing [11]. The building block of a CNN is a process called convolutions, unlike a neuron in a neural network, a neuron in a CNN layer is not connected to the entire input data, but to a specific window of the input image. Convolutional layers use sets of learnable filters (or Kernesl) slide them across the input data, and perform convolution with the input data [11]. The process of convolutions is done by multiplying each pixel by the corresponding filter value and summing the result to reduce a feature map. The idea behind this process is to allow the network to learn and capture local patterns and spatial information from a subset of the input data, which results in these feature maps capturing local patterns, edges, textures, lines [39], [23]. And because these learnable filters (weights) are shared, across the entire image, the network is translation invariant, meaning it can detect features regardless of their position in the input image, this allows the model to be more generalizable and invariant against different positions, and orientations of images [11].

After convolutional layers, the CNN network applies techniques to reduce the spatial dimensions of the feature maps, this is called the pooling layer. Pooling layers reduce the dimensions of the hidden layers by combining the outputs of the neurons at the previous layer into a single neuron in the next layer, depending on the pooling operation, the single neuron that represents the cluster of previous layers' neurons changes [7]. The pooling operation could be Max pooling, which is most widely used, this operation assigns the maximum value from each cluster of neurons to a single neuron at the next layer. Whereas average pooling calculates the average from the neuron clusters of the previous layer and assigns that average to the next layer [7]. Pooling layers make the network more robust against translations, and small variations in data, as it allows the model to reduce its sensitivity of exact feature positions in the input because where max pooling assigns the maximum value from a neighbouring cluster, this means that value would represent that region, making it more generalizable and less prone to overfitting [11], [23]. Finally, fully connected layers are used at the end which integrate the features learned through the preceding layers, after they have been downsized through the pooling layers, and are then used to make the final prediction, by mapping the features to a classification output. Depending on the number of classes specific to a downstream task, the classification can be adjusted [11].

Sparse connectivity and parameter sharing, as briefly mentioned, are among the primary reasons for CNNs' effectiveness in image processing. Unlike traditional neural networks, where each neuron is connected to every neuron in the next layer, leading to a large number of connections (parameters), specifically for higher resolution images, CNNs, have local receptive fields, meaning the network connects each neuron to only a small window of the input, this reduces computational complexity, and that is the idea behind sparse connectivity[11], [23]. Parameter sharing (filter sharing) refers to CNN's ability to share parameters across the entire input image, this is beneficial because it allows the model to detect the same features in different regions and parts of the image [11]. These attributes make CNN perform more effectively than traditional neural networks in terms of hierarchical feature learning, where the input representations get progressively more abstract as the layers increase [11]. Further, neural networks suffer from the dense connectivity that would arise from processing, resulting in the requirement of more computational power compared to CNNs. Further, CNNs tackle spatial locality through their convolutional filters, thereby making them more scalable. The advantages of CNN over traditional neural networks for large-scale image processing tasks were especially highlighted through AlexNet, a deep CNN that revolutionized large-scale image classification by introducing a highly optimized CNN architecture utilising deeper layers and regularization techniques that allow the model to be suited for large-scale processing [23].

2.2.2 Evolution of CNN Architectures

Convolutinal Neural networks has revolutionized the field of deep learning, and particularly computer vision, with that, its important to reflect on the evolution of the CNN architectures over time, and how they increasingly became more deeper, with each iteration advancing through addressing shortcomings of previous designs. This section will provide a basic overview of the chronological CNN evolution, from the early models like LeNet and AlexNet, which by modern standards are considered as shallow, to more recent and widely used architectures like ResNet.

2.2.2.1 LeNet

The first iteration in evolution of CNNs was LeNet-5, introduced by Lecun et al. in 1998, which was developed for hand-written digit recognition on the number MNIST dataset [24]. The architecture of LeNet-5 consisted of seven layers, including two convolutional layers, two pooling layers, and two fully connected layers [24]. One of the key principles introduced by LeNet-5 was the use of convolutional layers to learn local spatial information such as edges, and textures from the input images, and following this by the pooling layers which reduced dimensionality and computational complexity, leading to its high performance in the digit recognition. LeNet-5 showcased how local connectivity and parameter sharing can be used to handle image data in a way that could extract more robust information compared to traditional neural networks. Evenethough LeNet is considered a shallow architecture now, it laid the foundation for development of more complex CNN architectures.

2.2.2.2 AlexNet

AlexNet was yet another fundamental milestone in CNN development, introduced by Krizhevsky et al. in 2012 [23]. AlexNet offered a deep convolutional neural network that outperformed all the other models on the task of Large scale visual recognition using ImageNet dataset [23]. AlexNet's contribution was the introduction of multiple innovations that allowed CNNs to be used for large scale image processing tasks. The innovations included the usage of rectified Linear Units (ReLU) as the activation function, as compared to the traditional sigmoid and tanh activation functions, which saturate and make gradients to become increasingly small during backprobagation, to the point where such updates vanishes before it gets the chance to update the network.

$$ReLU(x) = max(0, x) \tag{2.1}$$

ReLU 2.1 on the other hand, helped the gradients to remain consistent by assigning 1 for positive values and gradient 0 for the negative values. Thus, allowing the gradients to remain consistent for the positive values, and in turn supports even deeper networks [23]. Besides using ReLU, AlexNet also used dropout as a regularization technique to promote model generalizability and reduce overfitting. Dropout refers to randomly dropping neurons during each training iteration. This allows the model to not rely on specifics of the learned features, allowing it to be more robust [23]. AlexNey consisted of five convolutional layers, max pooling in-between each convolutional layer, except between conv3, and conv4, and 3 fully connected layers, showing the performance of deeper CNN architecture on complex tasks such as Large scale image classification, and how to optimize such architectures [23].

2.2.2.3 Very Deep Convolutional Networks

Following AlexNet's success, VGG (Very Deep Convolutional Networks) was introduced by Simonyan and Zisserman in 2015 [44]. VGG introduced a CNN architecture that benefits from more convolutional layers, increasing the depth, and the use of smaller window sizes of 3x3, this allowed the network to go deeper, up to 19 layers, while still keeping the number of parameters manageable [44]. The use of small filter sizes, allowed VGG model to capture more complex and detailed features from the input data. While AlexNet used filter sizes of 11x11 and 5x5 in the initial layers, VGG was able to capture more details with the stacking of the small filters, leading to increasingly larger receptive fields, which acts as though a larger receptive field has been broken down [44]. The smaller window sizes, coupled with VGG's increased depth, enabled the architecture to capture even more fine-grained hierarchical features.

2.2.2.4 Residual Networks

The most advanced breakthrough in CNN architecture was with the introduction of Residual Networks (ResNet) introduced by He et al. in 2015 [14]. While earlier innovations like ReLU activation functions in AlexNet helped mitigate the vanishing gradient problem, they weren't sufficient for very deep networks, which were highly sought after. ResNet addressed this issue, as well as the problem of degradation in deep networks, through the introduction of skip connections, also known as residual connections [14].

The main innovation of ResNet is the residual block. In a residual block, the input x is passed through a series of layers to produce F(x) which represents the residual function, but x is also directly added to the output of these layers. This can be expressed as:

$$y = F(x) + x \tag{2.2}$$

Where F(x) in 2.2 represents the residual function that is learned by the stacked layers. This allows the network to learn residual mapping, while the skip connection preserves the input information. This architecture resolves issues like the degradation problem, where the performance of very deep networks deteriorates as the depth increases, despite lower training errors. The residual blocks also facilitate easier optimization, as the skip connections provide a direct path for gradients to flow backward through the network, preventing them from diminishing during backpropagation [14].

The skip connections allow the network to learn identity mappings directly. The idea behind residual learning is that the network layers focus on learning the residual, which is the difference between the input and the expected output, while the identity mapping ensures that information is preserved across layers. In ResNet, each residual block contains two or more convolutional layers, and the skip connection bypasses these layers, adding the input directly to the output. This is the core principle that allows ResNet to train very deep networks, enabling state-of-the-art performance in image classification[14]. Resent was implemented in several variants, including ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. The deeper variant, ResNet-152 reaches up to 152 layers, eight times deeper than the VGG network, all while maintaining relatively low computational complexity. ResNet-50 and deeper models also introduce a bottleneck architecture, which reduces the number of parameters by using 1×1 convolutions to downsample the input before applying the common 3×3 convolutions. [14].

2.2.3 Applications of CNNs in Medical Image Classification

Convolutional Neural Networks (CNNs) have gained significant attention in the medical imaging domain due to their ability to learn hierarchical representations of images and the benefits they offer through transfer learning. Their architecture is well-suited for handling spatial data, capturing patterns, textures, and fine-grained anomalies in medical images that often have much higher sensitivity than natural images [20]. One common issue in the medical domain is data sparsity, as annotated data is often limited, making the development of domain-specific models more challenging. In such cases, CNNs benefit from transfer learning, a technique that utilises pre-trained models on large-scale datasets, often natural image datasets such as ImageNet and fine-tunes them for medical tasks [20]. The success of transfer learning in medical image classification can be attributed to CNNs' ability to extract generic features from large-scale datasets, which are then fine-tuned for specific tasks with medical image datasets. Despite the differences between natural images and medical images, the early layers of CNNs are designed to extract low-level features such as edges, lines, and textures, which provide a strong foundation for adapting to medical imaging tasks. This ability to transfer learned knowledge is especially crucial in medical contexts, where data is often scarce [20].

Moreover, medical datasets not only suffer from limited data but are also often highly imbalanced, particularly in cases of disease diagnosis, where certain conditions or classes may be overrepresented. CNNs when combined with transfer learning, can mitigate these issues as as they are less sensitive to data imbalance and can generalize more effectively across underrepresented classes [52] [20].

2.2.3.1 Medical Image Classifications with CNN

The research paper titled **MedNet: Pre-trained Convolutional Neural Network Model for the Medical Imaging Tasks**[1] addresses the limitations of using general-purpose pre-trained models, like those trained on ImageNet, for medical imaging tasks. Medical images, such as MRI, CT scans, and X-rays, differ significantly from natural images, resulting in poor model performance when transfer learning (TL) is used. Furthermore, medical imaging datasets are typically small due to the time-consuming and costly manual labeling required by experts.

The authors propose MedNet, a CNN model specifically designed for medical imaging, aiming to bridge the gap left by existing general-purpose models. MedNet comes in two versions: Gray-MedNet, trained on 3 million grayscale medical images (MRI, CT, X-ray, ultrasound, PET), and Color-MedNet, trained on 3 million color images (histopathology, ophthalmology, etc.). This extensive training allows MedNet to learn features unique to medical images, enhancing its performance when fine-tuned on smaller datasets for specific medical tasks.

MedNet's architecture is designed to address common deep learning challenges, such as feature extraction, vanishing gradient [15], and overfitting. To train, the authors utilize data augmentation and generative adversarial networks [12] to manage data imbalance. After pre-training on large medical image datasets, MedNet can be fine-tuned on smaller, specific datasets to validate its effectiveness. This model offers several key benefits: it generalizes well across different medical imaging tasks, mitigates overfitting, reduces the need for extensive annotation, and is versatile enough for classification, segmentation, detection, and diagnosis.

The paper “**Classification of Alzheimer’s Disease Using fMRI Data and Deep Learning Convolutional Neural Networks**” [41] presents a deep learning approach to classify Alzheimer’s disease (AD) using functional Magnetic Resonance Imaging (fMRI) data. Alzheimer’s is a progressive neurological disorder, and its diagnosis often involves complex assessments. The study aims to develop a predictive model using deep learning, specifically CNNs, to differentiate between AD-affected brains and healthy brains.

The researchers sourced data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database, involving both elderly control subjects and AD patients. The fMRI data was preprocessed using techniques like motion correction and spatial smoothing to ensure high-quality input for the model. The preprocessed 4D fMRI data was then converted into stacks of 2D JPEG images suitable for CNN input.

The researchers adopted the LeNet-5 architecture, a well-known CNN model originally used for digit recognition, modifying it for binary classification of Alzheimer’s and normal fMRI data. The CNN architecture includes convolutional layers for feature extraction, pooling layers to reduce spatial dimensions, and fully connected layers for final classification. The convolutional layers play a crucial role in identifying high-level, discriminative features in the input images, such as patterns in brain structure and activity.

The results were impressive, with the model achieving a classification accuracy of 96.85%. This high accuracy illustrates the CNN’s strength in feature extraction and classification, outperforming traditional methods like Support Vector Machines (SVM).

Another application of CNNs within medical image classification is with the paper titled “**CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning**” [37]. This research paper introduces a deep learning model designed to automatically detect pneumonia from chest X-ray images with accuracy surpassing that of practicing radiologists. The model, called CheXNet, was trained on the largest publicly available chest X-ray dataset, ChestX-ray14, which contains over 100,000 frontal-view images labeled with 14 different diseases. The primary objective of this research is to create an algorithm capable of identifying pneumonia in chest X-rays while also being able to extend its application to other thoracic (chest) diseases.

CheXNet is a 121-layer CNN based on the DenseNet architecture. DenseNets enhance the flow of information and gradients throughout the network, making it possible to optimize such deep neural networks effectively. CheXNet takes a chest X-ray image as input and outputs the probability of pneumonia, along with a heatmap that highlights the areas of the image most indicative of the pathology. This interpretability is crucial in a medical context, providing visual evidence to support the model's predictions. The model is trained using a weighted binary cross-entropy loss function to handle the class imbalance present in the dataset, which contains a higher number of negative cases compared to positive ones.

To assess CheXNet's performance, the authors compared it against four practicing academic radiologists on a subset of 420 images. The evaluation metric used was the F1 score, which balances precision and recall. The results showed that CheXNet achieved an F1 score of 0.435, outperforming the average F1 score of 0.387 obtained by the radiologists. This difference was statistically significant, indicating that CheXNet not only matches but exceeds the diagnostic performance of radiologists on pneumonia detection. Additionally, when extended to detect all 14 diseases in the dataset, CheXNet outperformed previous state-of-the-art models on every class, marking a substantial advancement in automated medical imaging.

Despite its success, there are limitations in the study. Both the model and radiologists were restricted to using only frontal radiographs, even though lateral views are sometimes crucial for accurate diagnosis. Additionally, neither the model nor the radiologists had access to patient history, which could further inform the diagnostic process. These limitations suggest that while the model's performance is impressive, it may represent a conservative estimate of its potential under real-world clinical conditions.

2.3 Vision Transformers In Medical Imaging

2.3.1 Transformers

The transformer architecture was first introduced in the context of natural language processing (NLP) as a model capable of handling sequential data without the recurrent structure of traditional models like LSTMs. [47] At its core, the transformer operates on sequences of tokens, using a mechanism called self-attention to dynamically weigh the importance of each token in relation to others. This attention mechanism allows the model to capture long-range dependencies within the sequence by computing pairwise interactions between all token pairs. Transformers also include positional embeddings since, unlike CNNs or recurrent models, they have no inherent understanding of token order. The standard transformer model consists of alternating layers of multi-head self-attention and feedforward neural networks (MLPs), with residual connections and layer normalization for stable training. These features make transformers highly flexible and powerful for a variety of tasks, enabling their widespread success in NLP. However, applying this architecture directly to images is challenging, which the Vision Transformer (ViT) addresses by reinterpreting images as sequences of smaller patches.

2.3.2 Vision Transformers (ViT)

The Vision Transformer (ViT) model was introduced by Dosovitskiy et al. [10] in the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." This work challenged the dominance of CNNs in image recognition by proposing a new method using the transformer architecture. Traditionally, transformers rely on a sequence of tokens and self-attention mechanisms to process data, and they had not been widely adopted in vision due to the computational complexity of applying attention to images, which are large grids of pixels.

A key limitation of applying transformers to images lies in the enormous computational resources required. In an image, every pixel would need to "attend" to every other pixel, making the process quadratic in nature. For instance, even a relatively small 250x250 image would result in over 62.5 million pairwise interactions in a single attention layer, which is computationally prohibitive. While some approaches like local attention (used in CNNs) exist to address this issue by focusing on localized regions, ViT takes a different route by proposing global attention using image patches.

ViT divides an image into smaller, non-overlapping patches, such as 16x16 pixels. Each patch is flattened into a 1D vector and then linearly embedded to match the fixed input size required by the transformer. Unlike traditional CNNs, which focus on local pixel relationships through convolutional kernels, ViT processes the entire image globally by attending to all patches simultaneously. However, transformers naturally lack the concept of spatial positioning inherent in CNNs. To address this, ViT incorporates positional embeddings that encode the order of the patches. These embeddings are learnable parameters that inform the model about the spatial location of each patch, ensuring the model retains information about the image's overall structure.

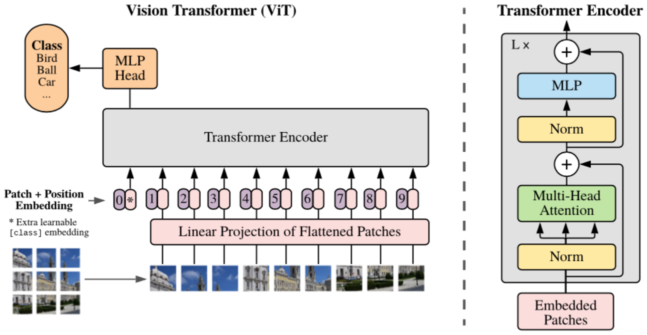


Figure 2.1: The ViT Architecture

ViT's architecture is similar to the standard transformer as seen in Figure 2.1

The primary advantage of ViTs is their ability to capture long-range dependencies from the outset, attending to the entire image globally rather than focusing on local pixel neighborhoods. This global attention allows the model to learn complex spatial relationships and patterns, which can be particularly beneficial when trained on large-scale datasets. While CNNs are designed with inductive biases like locality and translation equivariance, ViTs are more general-purpose and do not assume any inherent structure in the data, making them potentially more versatile when pre-trained on sufficiently large datasets.

One challenge in using transformers for vision tasks is the computational cost, especially for large images and high-resolution patches. ViT mitigates this by operating on relatively large patches (e.g., 16x16), reducing the sequence length and making global attention feasible. Despite this simplification, ViTs have shown remarkable performance, often surpassing CNNs in accuracy and training efficiency when pre-trained on vast datasets.

2.4 3D Medical Image Classification

Within medical imaging, 3D medical image classification is an essential component as it allows for the analysis of volumetric data to identify anatomical structures, tissues, or diseases. Unlike 2D images, 3D medical imaging offers comprehensive spatial information, allowing for a more accurate and detailed diagnosis. The main types of imaging modalities used in 3D classification include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Ultrasound. Each modality has its specific advantages: CT scans provide detailed images of bones and soft tissues, MRI is excellent for high-contrast soft tissue imaging, PET is valuable for functional imaging, especially in cancer detection, and Ultrasound is primarily used for imaging soft tissues in organs like the heart and liver.

Despite its promise, 3D medical image classification faces several challenges. The high dimensionality of 3D images leads to increased computational complexity and memory demands. Additionally, medical datasets often suffer from an imbalance between healthy and pathological samples, complicating the training process. Annotating 3D images requires expert knowledge and is labor-intensive, resulting in smaller labeled datasets.

2.4.1 3D CNNs

To tackle the challenges faced with 3D medical image classification, various deep learning techniques have been developed. The most commonly used are 3D Convolutional Neural Networks (3D CNNs), which extend traditional 2D CNNs to handle three-dimensional data by using 3D kernels that capture spatial relationships in volumetric images. Popular architectures include 3D U-Net [38], V-Net [32], and ResNet 3D. Due to the scarcity of large-scale annotated 3D medical datasets, transfer learning using pretrained 3D networks is frequently employed. Attention mechanisms are also used to help models focus on the most relevant regions in the 3D volumes, enhancing performance.

2.4.1.1 ACS Conv

While 3D CNNs have shown great promise in processing volumetric data, they come with significant limitations. The use of 3D kernels increases computational complexity and memory requirements, making model training and inference more resource-intensive. Additionally, the scarcity of large-scale annotated 3D medical datasets hinders the application of transfer learning, as there are few publicly available 3D pretrained models that are diverse and large enough to support universal 3D learning.

To address these challenges, *Axial-Coronal-Sagittal (ACS) convolutions* have been proposed as a novel solution in the paper titled **Reinventing 2D Convolutions for 3D Images** [50]. ACS convolutions reinvent the traditional 2D convolutions to operate on 3D data, allowing natively 3D representation learning while utilizing the pretrained weights from large 2D datasets like ImageNet. In ACS convolutions, the 2D convolution kernels are split by channels into three parts, each applied separately to the three orthogonal views (axial, coronal, and sagittal) of the 3D image. This design enables the model to capture comprehensive 3D spatial relationships using the efficiency and benefits of 2D convolutions.

A key advantage of ACS convolutions is that they allow any 2D CNN architecture (such as ResNet, DenseNet, or DeepLab) to be converted into a 3D model. This conversion provides an easy replacement for standard 3D convolutions, significantly reducing computational demands and model size while maintaining the ability to capture volumetric context. In addition, ACS convolutions benefit from the use of pretrained 2D weights, offering a significant performance boost in 3D medical image classification tasks, especially when annotated 3D datasets are limited.

2.4.2 3D Vision Transformers

3D Vision Transformers (ViTs) have recently gained traction in the field of medical image classification, particularly for 3D imaging modalities such as CT scans, MRI, and PET. Unlike traditional CNNs that excel at local spatial feature extraction, Vision Transformers use the Transformer architecture to process images, thereby capturing both global context and long-range dependencies. This global context awareness is crucial for interpreting complex 3D medical images where abnormalities or patterns might not be localized but spread across different parts of the volume.

One of the primary reasons for using Transformers in 3D medical imaging is their ability to handle large-scale volumetric data more efficiently. 3D medical images consist of numerous slices or volumes, leading to high-dimensional datasets. By breaking these volumes into smaller patches (tokens) and employing a self-attention mechanism, 3D Vision Transformers can encode intricate spatial relationships within the entire image. This capability significantly enhances the detection of subtle patterns, such as diffuse lesions in brain MRI scans, which are sometimes overlooked by CNNs focusing on local features [34].

The architecture of 3D Vision Transformers typically involves several key components. First, the 3D image is divided into a sequence of smaller, non-overlapping 3D patches (e.g., 16x16x16 voxels). Each patch is then flattened and embedded into a fixed-size vector, creating a series of tokens. To help the model understand the spatial relationships among these tokens, positional encodings are added. Unlike CNNs, which naturally process spatial information through convolutional filters, Transformers rely on these encodings to provide the necessary spatial context. The self-attention mechanism then processes each patch token in relation to all others, allowing the model to capture global features and long-range dependencies. Finally, multiple Transformer blocks, each comprising multi-head self-attention layers and feed-forward neural networks, are stacked to learn complex patterns within the 3D image. The representation of the patches is eventually pooled and passed through a classification head to output predictions, such as distinguishing between healthy and diseased tissue.

Recent developments in 3D Vision Transformers, such as Swin Transformers (shifted window-based Transformers) and MedFormer, have tailored the original Transformer architecture to handle 3D medical data more effectively. These models incorporate changes that enable more efficient processing of 3D volumes, thereby reducing computational overhead. Furthermore, pretrained Vision Transformers on large medical image datasets can be fine-tuned for specific tasks, improving their applicability in various clinical settings.

Using 3D Vision Transformers also comes with challenges. They have a high computational cost, especially when dealing with 3D volumetric data, which can be a limiting factor in resource-constrained settings. Additionally, Transformers typically require large amounts of labeled data to perform optimally, which is often a challenge in medical imaging due to the scarcity of annotated datasets. To address these issues, some research has focused on hybrid models [21] that combine the strengths of CNNs and Transformers, capturing both local features and global context. Transfer learning and self-supervised learning on large medical datasets also help mitigate data scarcity.

2.5 Foundation Models

Foundation models are large-scale AI models trained on large amounts of data to handle a diverse range of tasks. These models, which include language models like GPT-3 [3] and multimodal models that can process text, images, and other data forms, such as GPT-4 [35], BERT [6], CLIP [36], MedCLIP [48], serve as a base that can be fine-tuned for specific applications. They are characterized by their general-purpose nature and the capacity to be adapted across various domains without needing to be built from scratch each time.

Foundation models generally rely on self-supervised learning, allowing them to learn patterns and structures from unlabeled datasets. Once trained, foundation models can be fine-tuned on smaller, labeled datasets to perform specific tasks, such as language translation, image recognition, content generation, or in this dissertation's case, medical image classification. This approach not only speeds up the development process but also improves performance due to the extensive knowledge embedded in the model from its broad, foundational training.

2.5.1 Vision-Language Models

2.5.1.1 MedCLIP

MedCLIP: Contrastive Learning from Unpaired Medical Images and Text [48] is a novel vision-language model designed to tackle the unique challenges of medical image-text contrastive learning. It draws inspiration from the Contrastive Language-Image Pretraining (CLIP) model, which learns to match images and textual captions in natural image datasets. However, MedCLIP adapts this concept to the medical domain, where image-text datasets are significantly less abundant and more specialized. While CLIP leverages large-scale natural image-text pairs (400 million pairs), the scarcity of medical datasets with similar scale and granularity poses a significant obstacle for adopting such pretraining in medical applications. Moreover, medical images and their corresponding text have subtler and more fine-grained differences compared to natural images, making the direct application of CLIP's methodology insufficient.

The primary issues addressed by MedCLIP are data insufficiency and the specificity of medical semantics. Unlike natural image datasets, medical image-text datasets are limited in volume and variety, restricting the transferability and utility of models like CLIP in medical contexts. Additionally, medical imaging requires capturing intricate distinctions both in the images (such as different stages of a disease) and in the associated textual information (diagnostic reports). This necessitates methods that can handle these subtle differences effectively. Therefore, MedCLIP aims to address the data insufficiency issue and capture the nuanced medical meanings that are critical in this field.

MedCLIP tackles these problems in two primary ways. First, it introduces a novel approach by decoupling images and texts for contrastive learning, which significantly scales up the usable training data. This decoupling enables the model to handle sparse datasets by pairing images and texts based on their semantic meaning rather than relying solely on exact image-text pairs. Second, it eliminates false negatives by using medical knowledge to construct a soft semantic matching loss. Traditional contrastive learning methods in models like CLIP can mistakenly classify semantically similar medical images and reports as negative pairs, leading to inaccuracies. MedCLIP overcomes this by utilizing a soft semantic matching loss that leverages a semantic similarity matrix based on medical entities, ensuring that images and reports with similar medical meanings are correctly paired during training, even if they come from different patients.

The architecture of MedCLIP consists of three main components: knowledge extraction, vision and text encoders, and a semantic matching loss. The knowledge extraction module builds a semantic similarity matrix by extracting key medical entities from reports. This matrix is crucial as it serves as the foundation for pairing images and text. The vision encoder processes the medical images into embeddings, while the text encoder does the same for the extracted entities. The embeddings are then used to match images and text according to the semantic similarity matrix, ensuring that the pairs reflect medically relevant connections. The semantic matching loss further refines the model by training it to focus on meaningful associations between images and text, reducing the likelihood of false negatives.

MedCLIP's workflow involves processing medical images with labels and extracting key entities from the text. These extracted entities are then passed through the knowledge extractor to create the semantic similarity matrix, which facilitates the pairing of images and reports based on semantic meaning. Both the images and the corresponding reports are encoded into embeddings by the vision and text encoders, respectively. These embeddings are then matched using the semantic similarity matrix, guiding the training process to focus on medically relevant relationships between images and textual information.

A notable aspect of MedCLIP is its approach to text extraction. Instead of using raw sentences from reports, which can contain excessive or irrelevant information, MedCLIP extracts concise, key medical entities to represent the reports. This process ensures that only the most crucial aspects are captured, which simplifies the learning process. These extracted entities also serve as potential classes for the medical images, explaining why certain specific labels, such as "infectious," yield higher performance.

MedCLIP further improves its robustness and accuracy by employing an ensemble method. Since a single medical image can be characterized in multiple ways through different textual descriptions, the ensemble method generates five different sentence descriptions for each image. The model then calculates the similarity between the image and each of these sentences and averages these similarities to form the final ensemble prediction. This method enhances the model's capability to match images and text accurately, regardless of varying textual descriptions.

In terms of performance, MedCLIP surpasses state-of-the-art models in tasks such as zero-shot prediction, supervised classification, and image-text retrieval. Remarkably, MedCLIP achieves this with only 20,000 pre-training data points, underscoring its efficiency and effectiveness in addressing data scarcity. By decoupling images and texts, using a semantic matching loss, and incorporating medical knowledge, MedCLIP presents itself as a simple yet powerful framework for multimodal medical data analysis.

2.6 Related Work and Surveys

The work by Doerrich et al. [9] named "**Rethinking Model Prototyping through the MedMNIST+ Dataset Collection**" provides a comprehensive benchmark study using the MedMNIST+ dataset collection, expanding on the original MedMNIST v2 database to include higher resolutions (64x64, 128x128, and 224x224 pixels). Their systematic evaluation of both CNNs and ViT-based architectures addresses several key questions in medical image classification: the impact of input resolution, the necessity of compute-intensive architectures, and the optimal training schemes.

The authors benchmarked several well-established CNN architectures (VGG16, ResNet-18, DenseNet-121, EfficientNet-B4) alongside transformer-based models like ViT, CLIP, and DINO. These models were tested under varying resolutions and training paradigms, including end-to-end training, linear probing, and k-nearest neighbors (KNN) classification. Notably, the study reveals that CNNs continue to outperform ViT-based models in accuracy when using end-to-end training, suggesting that despite the recent surge in interest for transformer-based models, traditional convolutional architectures still have a competitive edge in medical image analysis tasks.

One observation from their work is the diminishing returns in performance observed with increasing input resolution beyond 128x128 pixels. Higher resolutions did not consistently improve model accuracy, challenging the prevalent assumption that higher-resolution inputs are necessary for reliable medical image classification. This finding advocates for using lower-resolution inputs, especially during the prototyping phase, to speed up processing while conserving computational resources.

The study emphasizes the potential of computationally efficient methods, such as k-NN classification integrated into the feature space of pre-trained models. By doing so, they highlight alternatives to resource-intensive end-to-end training, bridging the gap between computational cost and model performance. Interestingly, their analysis demonstrates that self-supervised pretraining strategies like CLIP and DINO do not necessarily lead to enhanced performance for end-to-end training but show improved results in linear probing and k-NN integration. These insights indicate that pretraining benefits can vary significantly based on the downstream training methodology, adding an important consideration for researchers working in medical image classification.

Huix et al. [16] take a different approach in their paper **Are Natural Domain Foundation Models Useful for Medical Image Classification?**, focusing on the transferability of state-of-the-art foundation models (SAM, SEEM, DINOv2, BLIP, OPENCLIP) to the medical imaging domain. The study addresses the key concern regarding domain shifts between natural images, on which most foundation models are trained, and medical images, which often exhibit more subtle, task-specific patterns. Given the ethical and privacy constraints in collecting large-scale medical datasets, transfer learning becomes a vital tool. However, adapting models pretrained on natural images to the medical domain remains challenging.

Their evaluation, conducted across four well-established medical imaging datasets (APTOS 2019, CBIS-DDSM, ISIC 2019, and CHEXPERT), shows that not all foundation models transfer well to medical image classification. Notably, DINOv2 outperforms the common practice of ImageNet pretraining, suggesting that its unsupervised pretraining strategy effectively captures semantic information useful for medical images. In contrast, models like SAM, BLIP, and OPENCLIP do not consistently outperform the ImageNet baseline, highlighting the necessity of domain-specific adaptation. The findings underscore that while low-level features in foundation models show some transferability, the higher-level features often require fine-tuning for optimal performance. This observation aligns with the view that freezing foundation models may lead to suboptimal outcomes, especially for complex medical classification tasks where high-level feature adaptation is crucial.

The study further delves into the architecture and training dynamics of foundation models, noting that stacking a classifier on top of foundation models results in only marginal performance gains. The study employs the Centered Kernel Alignment (CKA) metric to explore how learned representations change during fine-tuning. The results reveal that early layers of these models exhibit a higher degree of feature reuse, whereas deeper layers require substantial adaptation to suit medical tasks. This reinforces the importance of flexible fine-tuning approaches rather than freezing large portions of the model when dealing with medical imaging datasets.

2.7 Summary

This section provides a comprehensive overview of medical imaging and its integration with AI, focusing on the evolution and application of various deep learning techniques in medical image classification. It begins by discussing the diversity and complexity of medical imaging modalities, emphasizing their crucial role in healthcare diagnostics and the challenges associated with their interpretation. Deep learning is then introduced as a promising solution to aid in medical image analysis, addressing the growing demand for quick and accurate diagnostic tools.

A deeper dive is taken into deep learning and neural networks, highlighting how traditional machine learning approaches require manual feature extraction and how deep learning automates this process through hierarchical representation learning. This leads to the introduction of CNNs as a key innovation in medical image classification, describing their architecture, the principles behind their effectiveness, and their evolution through models like LeNet, AlexNet, VGG, and ResNet. The section also outlines CNNs' applications in medical image classification, underscoring their role in handling spatial data, capturing fine-grained details, and leveraging transfer learning for scenarios with limited annotated medical datasets.

Further, the section explores the emergence of ViTs in medical imaging, describing their architecture, ability to capture global spatial relationships, and application to 3D medical image classification. Challenges and limitations of these models are discussed, addressing models' computational costs and the need for large amounts of labeled data. Lastly, foundation models and their adaptation for medical image classification are explored, as well as the ongoing research on vision-language models like MedCLIP that address unique challenges in the medical domain.

3 METHODOLOGY

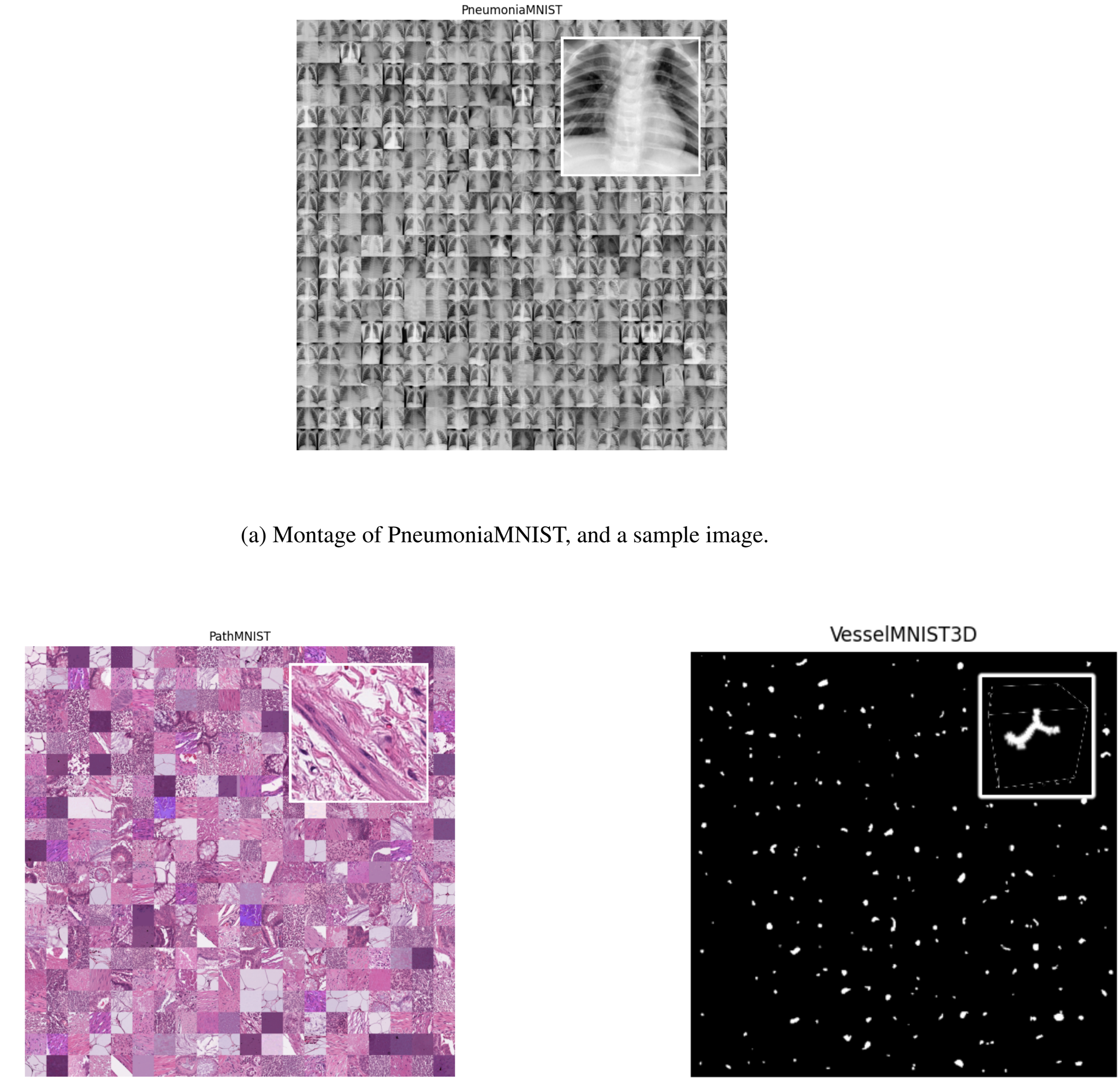
This chapter provides a detailed description of the methods used to evaluate and extend the work on medical image classification tasks using the MedMNIST dataset collection. The objective of this study is to assess how newer architectures, such as Vision Transformers(ViT), and vision-language foundation models like MedCLIP, perform on a variety of medical images with varying size, dimensions, and modalities, as compared to the more traditional approaches like CNNs as proposed by MedMNIST and set as the benchmark for the dataset collection. The methodology presented in this chapter begins with a detailed outline of the datasets used in this project, and is followed by an in-depth description of the various model architectures that were implemented, ranging from Resnet18 and ResNet50, used to reproduce the benchmark, to more advanced models like 3D ViTs and MedCLIP. Additionally, a comparison between MedCLIP and OpenAI's ChatGPT was conducted to assess the capabilities of vision-language models in zero-shot classification tasks. The final section details the training procedures, including hyperparameter choices, optimizer settings, and techniques used to handle imbalanced datasets. These training setups were applied across multiple models, with minimal changes between different datasets and models, to ensure consistency and provide a basis for comparing model performances and further experiments. By the end of this chapter, the reader will have a clear understanding of the technical setup of this work, allowing for the reproduction of experiments discussed in the following chapter.

3.1 Datasets

The datasets used in this project are part of the MedMNIST dataset collection, notably MedMNISTv2, and MedMNIST+, a lightweight benchmark designed for research and evaluation of machine learning models for biomedical image classification tasks [51]. The MedMNISTv2 collection consists of 12 2D and 6 3D datasets, each representing a unique medical imaging modality and classification task. The datasets are preprocessed into a standardized format—28×28 pixels for 2D images and 28×28×28 voxels for 3D images—allowing researchers to focus on model performance without the need for extensive preprocessing. In total, MedMNISTv2 includes 708,069 2D images and 9,998 3D images. The collection was developed to simplify and standardize medical image analysis, and has been developed carefully to address several key challenges in the field. MedMNIST includes a variety of data modalities from X-ray, MRI, pathology slides, OCT, and ultrasound, which helps in representing real-world medical imaging scnearious. This diversity further allows for research in the field to be more generalizable, as the diversity of the dataset modalities will allow research in the field to be on developing models on different types of medical data, and ensure generalizability across the different tasks and modalities [51], [46]. Further, the standardized format of 28x28 and 224x224 pixels for 2D datasets and 28x28x28 and 224x224x224 for 3D datasets eliminates the need for the extensive preprocessing and domain knowledge that has long been a struggle in the medical imaging field. With this, MedMNISTv2 and MedMNIST+ allows for simpler and more standardized approach for comparison between different models [51], [22]. Further, the images are preprocessed into an MNIST-like format, meaning a fixed, small resolution similar to teh original MNIST dataset. [CITE MNIST] which is important in the medical imaging domain, where preprocessing often requires domain-specific knowledge, such preprocessing could include adjustment for imaging modalities, scales, and organ-specific characteristics. MedMNIST eliminates this barrier by providing the datasets that are ready for immediate use, focusing on development of machine learning. Finally, MedMNIST was developed to serve as a benchmark dataset, similar to MNIST for digit classification. This allows researchers to assess and compare different machine learning models on a standardized platform, Baseline results for models like ResNet and AutoML were provided in the initial publication, which helps to establish a reference to future work.

3.1.1 Datasets Used

In this project, two 2D datasets and one 3D dataset from the MedMNIST collection were used: PathMNIST, PneumoniaMNIST, and VesselMNIST3D, to evaluate how models handle different image modalities, tasks, and dataset sizes. The two 2D datasets differ significantly in their classification tasks, sample sizes, modalities, and image formats. To further explore the impact of image resolution, I worked with both the 28×28 images from MedMNIST v2 and the larger 224×224 images from MedMNIST+. PathMNIST is a multi-class classification task with the largest sample size, with 100,000 non-overlapping image patches from hematoxylin & eosin stained histological images, while PneumoniaMNIST is a much smaller, binary classification task on 5,856 pediatric chest X-Ray images. VesselMNIST3D, a 3D dataset with a binary classification task, obtained based on an open-access 3D intracranial aneurysm dataset, IntrA34, containing 103 3D models (meshes) of entire brain vessels collected by reconstructing MRA images, which MedMNIST generated 1,694 healthy vessel segments and 215 aneurysm segments from the models. VesselMNIST3D is even smaller and has significant data imbalance. This deliberate selection of diverse datasets was designed to test how well different architectures generalize across varied medical imaging tasks. My goal was to push the models to perform in different settings, making the evaluation more robust and reflective of real-world applications. Figure 3.1 below shows a montage of the three selected datasets, as well as a sample image from each set. Table 3.1 shows necessary information on each datasets, such as modality, the classification task, image sizes considered, and their number of samples.



(a) Montage of PneumoniaMNIST, and a sample image.

(b) Montage of PathMNIST, and a sample image.

(c) Montage of VesselMNIST3D, and a sample image.

Figure 3.1: Visual representation of the chosen datasets

Table 3.1: The three chosen datasets from MedMNIST collection. MC: Multi-Class. BC: Binary-Class. v2: MedMNISTv2. +: MedMNIST+

Dataset	Modality	Task	Image Size	# Samples
PathMNIST	Colon pathology (H&E)	MC (9)	28×28 (v2) 224x224 (+)	107,180
PneumoniaMNIST	Chest X-ray (CXR)	BC	28×28 (v2) 224x224 (+)	5,856
VesselMNIST3D	Brain MRA	BC	28×28x28(v2)	1,908

The datasets used were already split to train, validation, and test sets, with ratio of 9:1 for PathMNIST. For PneumoniaMNIST, ratio of 9:1 into training and validation set, and the dataset's source validation set has been used as the test set. Finally, ratio of 7:1:2 into training, validation and test set has been set for VesselMNIST3D.

3.2 Data Preprocessing

The MedMNIST dataset collection is designed to be standardized, making it accessible without the need for domain-specific or other preprocessing. This project, therefore, aimed to maintain alignment with the preprocessing steps taken by the MedMNIST team for their benchmarking to ensure a fair comparison of model performance. The following subsection will describe the preprocessing steps used by MedMNIST for the dataset collection and their benchmark models. Then, this project's preprocessing will be described on a general level, followed by model-specific preprocessing, with justifications provided for any deviations from MedMNIST's approach. Experiment-specific preprocessing will be outlined in the following chapter on experimental results.

3.2.0.1 MedMNIST Preprocessing

To ensure the MedMNIST datasets are standardized and lightweight, the dataset images from MedMNISTv2 are in the preprocessed MNIST-like format of 28×28 pixels for 2D images and 28×28×28 voxels for 3D images, as discussed above in section 3.1. This resizing of diverse medical images from different sources and modalities simplifies the input requirements for various models, ensuring a uniform input size, and thereby making it easier to test different models quickly and benchmark performance across different datasets and models. The datasets undergo cubic spline interpolation to scale the images down to the target 28x28 and 28x28x28 in MedMNISTv2. This process allows for scaling down the images while preserving as much of the image's detail and structure as possible. When it comes to splitting, MedMNIST provides official train-validation-test splits to avoid data leakage and ensure fair benchmarking, as discussed above in section 3.1. With this official splitting, MedMNIST ensures that researchers can focus on evaluating model performance without concerns about such preprocessing challenges.

Regarding MedMNIST's benchmarked models, which include ResNet architectures, AutoML tools like auto-sklearn, AutoKeras, and Google AutoML Vision (a commercial AutoML solution), they applied the following preprocessing techniques. These will be closely followed and aligned with for this project's experiments to ensure fair comparisons:

- Normalization of mean and standard deviation of 0.5 applied to the images.
- Images from datasets containing greyscale images like PneumoniaMNIST are converted to RGB by replicating the single channel three times. This was done to ensure compatibility with model architectures such as ResNet and ViTs.
- MedMNIST utilizes resizing techniques for model architectures that require higher-resolution inputs, such as larger ResNets or ViTs. The resize option allows for resizing 28x28 2D images to 224x224.
- For 3D datasets, the images are voxelized into 28x28x28 format.
- For the 3D shape modality datasets, a form of data augmentation is applied to some 3D datasets, including VesselMNIST3D, which involves random scaling of voxel values during training. In this process, voxel values are multiplied by a random value between 0 and 1 to introduce variability in the input data, allowing the model to generalize better. During evaluation, the voxel values are scaled by a fixed coefficient of 0.5.

3.2.0.2 Preprocessing for the Models

In this project, the goal was to align with MedMNIST's preprocessing pipeline as much as possible to ensure that any improvements in performance are directly attributable to t he models and not the preprocessing. As such, the same preprocessing steps of normalization, color conversion, and

3D voxel scaling, as mentioned in section 3.2.0.1, were followed as much as possible. However, since the objective of this project was to test the performance of new and different models, slight preprocessing adjustments were made to align with model requirements at times. In summary, the preprocessing steps taken throughout this project were as follows:

- Image Resizing: While the MedMNISTv2 datasets provide 28×28 images for 2D tasks, models with higher-resolution input requirements were explored, such as Vision Transformers (ViTs) and Vision-Language foundation models with ViT backbones, which require larger input sizes. In those cases, the 28x28 images were resized to 224×224 pixels to match the input requirements of ViTs, MedViT, and MedCLIP, using the same resizing function provided by MedMNIST to ensure the resizing technique adhered to MedMNIST's preprocessing steps. However, for some model experiments, an image size of 224x224 was used directly from the MedMNIST+ collection to assess model performance. In such cases, images were not resized and were directly compared with MedMNIST's 224 image size benchmarks.
- Grayscale to RGB Conversion: Similar to MedMNIST's practice, grayscale images from datasets like PneumoniaMNIST were converted to RGB by replicating the single channel three times. This ensures compatibility with models like ResNet and ViTs, which expect three-channel input.
- Normalization: Consistent with MedMNIST's standardization, all images were normalized to have a mean of 0.5 and a standard deviation of 0.5, aligning directly with MedMNIST's pipeline to maintain uniformity in the input data across experiments.

Model-specific preprocessing was as follows:

MedViT

- For MedViT, besides resizing 2D images to 224x224 using MedMNIST's resizing function, AugMix was applied as a data augmentation technique to introduce variability in the training data. This augmentation was implemented to align with the preprocessing steps and guidelines set by MedViT in their implementation.

3D Models

- The 3D dataset VesselMNIST3D was voxelized using a custom Transform3D class that converts voxel data into tensors. This was followed by normalization of the voxel values, which were also scaled during training and evaluation, following MedMNIST's steps for handling 3D shape modality datasets.
- Experiment-specific preprocessing was applied to VesselMNIST3D, which will be explored in detail in the following chapter. These preprocessing steps included a variety of data augmentations to aid in model generalizability due to the dataset's high class imbalance, which was more apparent in the ViT model performance compared to the ResNet models used in the benchmark.

MedCLIP

- For the MedCLIP zero-shot prompt-based classification, besides image resizing, preprocessing was mainly handled by the CLIP processor, which includes automatic resizing and normalization of the input images to match the expected model input.

No additional preprocessing was applied to ensure that the pipeline remained in line with MedMNIST's principles, only the standard resizing and tensor conversion. This approach allows for fair comparisons with the MedMNIST benchmarks, ensuring that any performance improvements come from the models rather than the preprocessing steps.

3.3 Model Architectures

This section details the various model architectures employed in this project for biomedical image classification using the MedMNISTv2 and MedMNIST+ dataset collection. Fine-grained details of each architecture are discussed in the literature review in Chapter 2. In this section, a high-level explanation for each architecture is provided, alongside the rationale behind choosing each model, and how they were implemented and adapted for this project. The project began with reproducing the benchmark models, including ResNet18, ResNet50, and ResNet18 with Axial-Coronal-Sagittal (ACS) convolutions for VesselMNIST3D. This was done to establish baseline model performance, and reproducing the benchmark models provided a deeper understanding of how MedMNIST's benchmarked models were implemented and evaluated. Subsequently, Vision Transformers (ViTs) were explored, testing different pretrained variants to identify the most suitable one, followed by developing a custom 3D ViT model architecture for VesselMNIST3D.

Advanced ViT architectures pretrained on medical images, such as MedViT, and Vision-Language foundation models like MedCLIP, were then employed to evaluate their effectiveness on different medical image classification tasks and datasets. The following subsections describe these architectures, their configurations, and how they were adapted for this project.

3.3.1 Benchmark Model Architectures

First, the benchmark models were reproduced to establish a baseline for this project. Since the aim was to replicate the results, MedMNIST's methodology and training setup were closely followed. Through experiments and informed decisions, the benchmark model results were successfully reproduced. For this, ResNet18 and ResNet50 for the 2D datasets were reproduced, and ResNet18 with ACS convolutions was implemented for the 3D dataset VesselMNIST3D. The following steps outline the architectural details adapted from MedMNIST to reproduce the benchmarks.

3.3.1.1 ResNet18 and ResNet50

The ResNet architectures utilized in this project to reproduce the benchmark models follow the original design of ResNet, using skip connections to mitigate the vanishing gradient problem, thus enabling the training of deeper networks. ResNet18 and ResNet50 differ in their block designs, with ResNet18 using four sequential stages of BasicBlock modules, while ResNet50 uses Bottleneck blocks.

- ResNet18: This model consists of an initial 3x3 convolutional layer, which is followed by batch normalization. The 3x3 convolution allows for capturing smaller details in the first layer. The core of the model consists of four stages, each containing a series of BasicBlocks. Each block includes two 3x3 convolutional layers, batch normalization, and ReLU activation. The skip connections in the blocks help stabilize the training process. The network ends with an adaptive average pooling layer and a fully connected layer responsible for classification. The adaptive average pooling before the fully connected layer allows the model to handle variable input sizes.

- ResNet50: This deeper ResNet variant uses Bottleneck blocks, which have three layers per block (1x1, 3x3, and another 1x1 convolution). This design allows the model to learn more complex features by increasing the depth of the network.

The input size and number of channels were adjusted according to the dataset. For PneumoniaMNIST, to ensure alignment with MedMNIST's implementation, the grayscale images were converted to RGB images as outlined in Section 3.2.0.2. The class number reflected the binary classification task for the dataset, while for PathMNIST, the number of classes reflected the multi-class task of the dataset, which is a 9-class classification.

3.3.1.2 ResNet18 + ACS Convolution for 3D Data

The standard ResNet18 model was modified to include Axial-Coronal-Sagittal (ACS) convolutions for handling 3D data, as implemented by MedMNIST. ACS convolutions allow 2D CNNs to perform native 3D representation learning by splitting 2D kernels into three parts for axial, coronal, and sagittal views [50]. This approach bridges the gap between 2D and 3D convolutions, making it possible to use pretrained weights from 2D datasets. ACS convolutions facilitate the processing of volumetric medical image data, such as VesselMNIST3D, using pretrained 2D weights. This is beneficial given the scarcity of large-scale, annotated 3D medical image datasets. Additionally, the implementation of ACS convolutions is simple, providing yet another advantage for its usage. By using ACS convolutions, the network maintains a smaller model size and less computational complexity than traditional 3D CNNs, which was the reason for their adoption by MedMNIST to produce their benchmarks.

3.3.2 Vision Transformer (ViT) Architectures

The Vision Transformer (ViT) architecture was explored for both 2D and 3D datasets in this project to assess its performance on medical image classification. The choice of ViT models, implementation details, and rationale for specific configurations are discussed in this section.

3.3.2.1 ViT for 2D Datasets: PneumoniaMNIST and PathMNIST

ViTs were fine-tuned for the 2D datasets, PneumoniaMNIST and PathMNIST to explore the potential of transformer-based models in biomedical image classification. The variant used was Vit-Base with patch size of 16 (ViT_B/16), this variant refers to the ViT with patch size of 16, and the Base model is one of the three sizes used for the ViT (Base, Large, and Huge). The Base refers to the configuration of the model that is directly adopted from the BERT configuration. The three configurations differ in terms of number of layers, hidden size, and other architectural parameters. Notably, the Base model in ViT configuration comprises of 12 layers, with hidden size 768, and MLP size of 3072, and uses 12 attention heads, the details of the ViT architecture are discussed in the literature section above. The ViT_B/16 has a good balance between model complexity and performance, as this variant has shown strong results across various tasks including more fine-grained tasks, similar to biomedical image classifications, while still being more manageable in terms of computational resources as compared to the larger variants [42]. Further, the patch size of 16x16 was an appropriate size for capturing features without loosing too much spatial information as compared to larger patch sizes. These advantages of ViT_B/16 pair with the aim of this project, where model architectures are intended to be assessed and experimented with, resulting in many experiments over the datasets, for this, a lighter variant of ViT that captures the core principles of vision transformers without the burdens of heavier computational cost, made this variant a suitable choice for the ViT model experiments.

Three ViT_B/16 models were tested from different sources, ViT_B/16 from HuggingFace, pretrained on ImageNet21K and fine-tuned on ImageNet-1k, and ViT_B/16 from PyTorch Image Models library (timm) library pretrained on ImageNet21k. Each implementation was tested by running inference and then fine-tuning for two epochs only to determine the better-suited implementation to be used in this project. The hyperparameter settings chosen for this fine-tuning test was designed to align closely to MedMNIST's hyperparameter settings while still respecting the requirements of a ViT-based architecture. The hyperparameters testing the ViT models with included learning rate of 1e-4, batch size of 64, and 2 epochs. Table 3.2 shows the result of running inference and fine tuning of PneumoniaMNIST on both ViT_B/16 implementation and their performance.

Table 3.2: Performance comparison on two different implementations of ViT_B/16 variants on PneumoniaMNIST done for the purpose of choosing a suitable ViT_B/16 variant implementation. HF: Hugging Face

	Inference		Fine-Tune	
	AUC	ACC	AUC	ACC
HF's ViT_B/16	0.212	0.375	0.812	0.789
Timm's ViT_B/16	0.494	0.416	0.947	0.876

The timm library's implementation of ViT-B/16 is better suited for this project's biomedical image classification tasks on PathMNIST and PneumoniaMNIST datasets as shown in 3.2 for several reasons. Timm's version offers advanced training techniques like AugReg (Augmentation and Regularization), which includes stronger data augmentation, stochastic depth, and mixup/cutmix regularization. These features are specially beneficial for medical imaging tasks where data can be limited or imbalanced. For PathMNIST, a large dataset for 9-class classification, timm's robust augmentation and regularization techniques can help prevent overfitting and improve generalization. For PneumoniaMNIST, a smaller dataset for binary classification, timm's ViT_B/16 highlights this ViT's variants' handling of smaller datasets and transfer learning capabilities.

The ViT_B/16 model from timm's library was then selected and used for the fine-tuning experiments on the two 2D datasets. For this, the ViT model was created, and the pretrained weights were loaded. The final classification head was then modified: the original fully connected layer was replaced with a new linear layer that maps the output features to the required number of classes in each dataset (binary for PneumoniaMNIST and multi-class for PathMNIST). To align with MedMNIST's preprocessing techniques while accommodating input size requirements for ViT architectures, image sizes of 28x28 were resized to 224x224 following the techniques mentioned in the preprocessing section 3.2.0.1. However, for some experiments, images of size 224x224 were directly used from the MedMNIST+ collection.

3.3.2.2 Custom ViT for 3D Datasets

For the 3D dataset VesselMNIST3D, a custom 3D ViT was implemented due to the incompatibility of standard ViTs with 3D volumetric data. Standard ViTs are designed for 2D images represented as 3D tensors (height, width, channels). In contrast, 3D medical volumes are 4D tensors (depth, height, width, channels). This introduces unique challenges, as directly adapting 2D ViTs to 3D data would significantly increase computational complexity and memory requirements. Additionally, 3D images contain volumetric spatial dependencies, requiring a model that can capture relationships across all three spatial dimensions effectively. These challenges and differences inspired the development of a custom 3D ViT, designed to handle 3D input by using 3D patch embeddings, modifying the positional encoding, and adapting the transformer architecture to process 3D volumes.

Key aspects of the 3D ViT implementation

- 3D Patch Embedding: The model first divides the 3D input volume of 28x28x28 voxels into smaller, non-overlapping 3D patches of size 7x7x7. These patches are then flattened into vectors, and a linear layer projects these flattened patches into a higher-dimensional embedding space (256 dimensions). This embedding class was custom-designed to create patches from 3D data, unlike traditional ViTs, which only operate on 2D patches. The class token is also added to the sequence of patch embeddings. This token serves as a global representation that aggregates information from all patches during the transformer's self-attention mechanism. Positional embeddings are then added to both the patch embeddings and the class token to retain spatial information in the volumetric space.

- Transformer Encoder: The transformer encoder in the 3D ViT consists of six layers, each containing multi-head self-attention (with 8 heads) and feed-forward neural networks. The embedding dimension (256) allows the model to effectively represent complex spatial relationships within the volumetric data. The encoder captures interdependencies between the 3D patches, allowing the model to learn volumetric features across depth, height, and width.

- Classification Head: After passing the embeddings through the transformer encoder, including the class token, global average pooling is applied to aggregate the learned features across all patches. The aggregated feature vector, along with the class token, is then passed through a multi-layer perceptron (MLP) head with layer normalization for final classification. The class token, in particular, acts as a global descriptor of the input volume, helping in the final decision making process.

For the ViT 3D implementations, ACS convolutions were not used as in MedMNIST's benchmark ResNet. This is because the custom ViT model was designed to inherently handle volumetric data through the described patch embedding mechanism, which divides the 3D volume into smaller patches. This approach directly learns spatial relationships in the voxel space without the need for specialized convolutions. Furthermore, the aim was to maintain the pure attention-based approach of ViT. The transformer mechanism processes spatial information through self-attention rather than localized convolutions as done through ACS convolutions. The goal here was to assess the performance of a fully transformer-based architecture on 3D data without incorporating convolutions; thus, the use of ACS convolutions for the 3D ViT model was omitted.

3.3.3 MedViT

MedViT was explored for fine-tuning on the 2D datasets PathMNIST and PneumoniaMNIST because of its unique design tailored for medical imaging tasks. Unlike standard Vision Transformers

(ViTs), MedViT is a hybrid model that combines CNNs with Transformers, addressing the need for both local texture recognition and global feature extraction in medical images, as CNNs due to their inductive bias, are more efficient in capturing spatial information, while the self-attention cabalities of transformers would allow for global feature extraction and context based understanding. The architectural modifications enable MedViT to generalize across different medical image domains, it was pre-trained on a diverse range of medical imaging data, including CT scans, X-rays, ultrasounds, and OCT images. This pre-training provides a strong foundation for the binary and multi-class classifications of PathMNIST and PneumoniaMNIST.

For the implementation in this project, the MedViT_small variant was chosen from the MedViT repository, because it balances model complexity with computational efficiency. The model was initialized with pre-trained weights designed for medical data. The original classification head of the model was modified to match the number of classes in each dataset, adjusting for the binary classification in PneumoniaMNIST and the multi-class task in PathMNIST. Images were resized to 224x224 pixels using MedMNIST's preprocessing techniques to accommodate the input size requirements of the MedViT architecture. The model was then fine-tuned using an SGD optimizer with a learning rate of 0.005 and a batch size of 10 for 10 epochs. These hyperparameter settings were recommended by MedViT to start with fine-tuning, other hyperparameters more inlined with the other ViT models used have been experimented with as well. To boost model robustness, as recommended by MedViT, data augmentation techniques like AugMix were used during training.

3.3.4 Joint 3D Architecture ResNet18 + ViT3D

The motivation for this joint architecture stems from the observed performance gap between ResNet18 with ACS convolutions (Benchmarked by MedMNIST and reproduced in this project) and the ViT3D model on the VesselMNIST3D dataset. While ResNet18 + ACS Convolutions provided the highest performance, even with the highly imbalanced VesselMNIST3D dataset, ViT3D struggled to achieve better results. The idea behind this hybrid model was to leverage the strengths of ResNet18 with ACS convolutions in extracting low-level features, while using the ViT3D's self-attention mechanism to capture long-range dependencies within the 3D volumes. By combining these two architectures, the objective was to improve the overall classification performance by providing the network with a more complex and high level set of features.

In this joint architecture, the ResNet3D-ViT3D model is composed of two primary components: a ResNet18 model adapted for 3D inputs and a custom 3D Vision Transformer. The

ResNet18 model was adapted to 3D data by modifying the first convolutional layer to a 3D convolution and converting all BatchNorm2d layers to BatchNorm3d. This allows ResNet18 to process 3D input volumes and learn spatial features along the depth, height, and width of the voxel data. Additionally, to align with the benchmark setup provided by MedMNIST, the ACSConverter was applied to this ResNet18, replacing the standard 3D convolutions with ACS convolutions, capturing information across different planes of the 3D volume.

The custom ViT3D model was designed to operate alongside the ResNet18. The implementation details for the ViT3D model was discussed in section 3.3.2.2 above. In this hybrid model, the output features from both the ResNet18 and ViT3D are concatenated into a single feature vector. A fully connected layer then combines these features and maps them to the final output class. During training and evaluation, this joint model is initialized with both the ResNet18 and ViT3D components. The input data is processed through the modified ResNet18 first, followed by the ViT3D. The outputs of these two components are concatenated to form a combined feature vector. This vector is passed through a linear layer for final classification. The use of both the ACS-converted ResNet18 and the ViT3D was with the intention to exploit their complementary feature extraction capabilities.

3.3.5 MedCLIP Fine-Tuning and Zero-Shot Prompt-Based Classification

MedCLIP was utilized in two ways in this project. Unlike general CLIP models, which are pretrained on internet image-text pairs, MedCLIP uses medical images and clinical reports for training, addressing the challenges in medical image classification by decoupling images and text for multimodal contrastive learning, eliminating false negatives in its learning process, and allowing training on various types of medical data, including unpaired images and text [48]. These are the reasons for choosing this model architecture for this project. The backbone of MedCLIP includes a visual encoder based on a vision transformer (ViT), enabling it to process complex spatial information in medical images.

3.3.5.1 Fine-Tuning MedCLIP

The rationale behind fine-tuning MedCLIP is to utilize its existing knowledge from being pretrained on medical images and reports. In this project, the vision-only MedCLIP model with a ViT backbone was fine-tuned on the MedMNISTv2 datasets (PneumoniaMNIST and PathMNIST) to evaluate its performance. MedCLIP's architecture allows it to encode images into embeddings through a ViT encoder. For the fine-tuning process, the model's visual encoder was initialized using the MedCLIPVisionModelViT class. This encoder processes images, generating feature embeddings which were then passed through a custom classification head. The classification head, a linear layer of 512 output units, was added to align with the number of classes in the MedMNISTv2 datasets. Pretrained weights were loaded into the ViT encoder to utilize MedCLIP's feature representations.

3.3.5.2 Zero-Shot Prompt-Based Classification using MedCLIP

The zero-shot prompt-based classification capabilities of MedCLIP were explored to assess the feasibility of using vision-language models for image-only datasets like MedMNISTv2. The idea here was to leverage MedCLIP's pretrained knowledge and its ability to align textual descriptions with visual patterns in medical images. This method bypasses the need for retraining or fine-tuning, offering a fast and flexible classification approach that is particularly usefull for datasets without associated textual data.

In implementing zero-shot classification, custom text prompts were created to represent each class label. For the PneumoniaMNIST dataset, prompts included descriptions like "This is a normal chest X-ray" and "This X-ray shows signs of pneumonia." MedCLIP's processor was used to convert these text prompts into embeddings, while the image inputs were similarly processed into visual embeddings by MedCLIP's ViT-based visual encoder. The model then calculated similarity scores between the image embeddings and each of the textual prompts, ultimately predicting the class with the highest similarity score. This approach allowed for testing MedCLIP's alignment of visual features with descriptive text without requiring any further training.

The experiments were not limited to simply applying MedCLIP's zero-shot capabilities. Different sets of prompts were tested to identify how the specificity and nature of the textual input impacted the model's performance. For instance, more descriptive prompts such as "a photo of a chest X-ray showing signs of infection such as infiltrates or consolidation" were compared against simpler prompts like "a photo that is normal." In some instances, I cross-referenced MedCLIP's results with GPT-4, examining whether the mistakes made by MedCLIP were consistent with the errors that a language model might make when interpreting similar visual cues. This cross-referencing was done to provide a deeper understanding of the limitations and strengths of using language models like MedCLIP for visual classification tasks. The detailed findings from this comparison are discussed further in the experimental results section.

For the PathMNIST dataset, which contains more complex and varied pathology classes, descriptive prompts were generated to capture the unique visual characteristics of each tissue type. The zero-shot approach enabled the model to process these medical images in an intuitive, flexible manner without the need for traditional training. However, the results varied depending on the prompts used, revealing the model's sensitivity to the specificity and quality of the input text.

3.4 Training Setup

For this project, the training setup aimed to closely follow MedMNIST's benchmark methodology to ensure fair and accurate comparisons. Additionally, adjustments to hyperparameters were made to accommodate the different model architectures, particularly for Vision Transformers (ViTs) and 3D datasets, as well as to explore the performance implications of different training configurations. This section outlines the training settings employed for the reproduction of benchmark models, fine-tuning ViT models, and training MedViT and MedCLIP.

3.4.1 Benchmark Model Training

To reproduce the benchmark results on MedMNISTv2 datasets, the same hyperparameters used by MedMNIST were employed. For the 2D datasets PathMNIST, and PneumoniaMNIST, the models were trained for 100 epochs using a batch size of 128 and an initial learning rate of 0.001. A weight decay was applied to reduce the learning rate by a factor of 0.1 at epochs 50 and 75. For the 3D dataset VesselMNIST3D, a smaller batch size of 32 was used due to the higher computational requirements for 3D data. Similarly, the learning rate and the epoch milestones for weight decay followed the same pattern as in the 2D setup.

The training employed the cross-entropy loss function, with the Adam optimizer to optimize the model parameters. This training procedure was standardized across ResNet18, ResNet50, and the ResNet18 withACS convolutions, using the MedMNISTv2 dataset classes to ensure a consistent comparison between these benchmark models.

3.4.2 Vision Transformer Training

When training ViT-based models, hyperparameter adjustments were necessary to account for the architectural differences between transformers and traditional CNNs. Notably, the learning rate was decreased to 0.0001, as ViTs generally benefit from lower learning rates during training [45]. Additionally, the batch size was reduced to 64, as the ViT architectures have higher memory requirements, especially when working with high-dimensional inputs, therefore, for time and resource management, the batch size was lowered, specially since the learning rate was decreased, a lower batch size was better suited with the smaller learning rate. The number of training epochs was set to 20 and 50 in separate experiments to explore how the ViTs perform under different training durations.

For the 3D ViT models, further adjustments were made to accommodate the increased computational complexity of 3D data. The batch size was further reduced to 8 for many experiments to manage memory usage during training. The learning rate of 0.0001 was maintained to facilitate the effective training of these deeper transformer architectures on volumetric data. These changes were motivated by the need to balance between computational feasibility and model performance.

Table 3.3: Project Baseline Hyperparameters set for model experiments

Baseline ViT Hyperparameters	Epochs	Batch Size	Learning Rate
ViT based 2D models	50/20	64	0.0001
ViT3D model experiments	20	8	0.0001

3.4.3 MedCLIP Fine-Tuning Training Setup

The MedCLIP model, which uses a ViT backbone, was fine-tuned for 20 epochs using a batch size of 64 and an initial learning rate of 0.0001. The choice of a lower learning rate and a reduced batch size was to accommodate the pre-trained nature of the MedCLIP model and to mitigate potential overfitting. Cross-entropy loss was utilized to optimize the model during training. Since MedCLIP has been pre-trained on medical image-text pairs, this fine-tuning setup aimed to assess how well the model could transfer its learned representations to new classification tasks within the MedMNISTv2 dataset collection.

3.4.4 MedViT Training

For training the MedViT models, the setup was slightly modified to align with the original MedViT paper. The models were trained for 10 epochs with a batch size of 10, however, experiments were done to match the consistent hyperparameter settings selected for all the ViT based models. The learning rate was set to 0.005, and the optimizer used was Stochastic Gradient Descent (SGD) with a momentum of 0.9. Weight decay was applied, with the learning rate adjusted by a factor of 0.1

at epochs 10 and 15, accommodating the model's complexity while preventing overfitting. This training setup aimed to use the MedViT's pre-training on medical images for effective fine-tuning on specific tasks within the MedMNISTv2 datasets.

The zero-shot experiments with MedCLIP did not involve a traditional training process. Instead, these experiments utilized MedCLIP's pre-trained vision-language capabilities to classify medical images based on text prompts. Since no model weights were updated during this phase, there is no specific training setup to report for this approach. However, the findings and the exploration of prompt engineering's impact on classification results will be detailed in the experimental results section.

3.5 Evaluation Metrics

To maintain strict alignment with MedMNIST's evaluation standards, same metrics were adopted as MedMNIST's original benchmarking: Area Under the ROC Curve (AUC) and Accuracy (ACC). By implementing these metrics consistently across all models, comparability of performance between various models, architectures, and training configurations were ensured. However, in the case of the Zero-Shot Prompt-Based Classification with MedCLIP, only accuracy was used since AUC requires model training and a continuous prediction score, which is not applicable in a zero-shot scenario.

3.5.1 AUC and Accuracy Metrics

AUC (Area Under the ROC Curve) is a metric for evaluating the performance of binary and multi-class classifiers. It represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, and the AUC value quantifies the overall separability of the model's predictions [13].

For binary classification, the AUC can be defined as:

$$AUC = \int_0^1 TPR(FPR^-1(x))\,dx \tag{3.1}$$

AUC ranges from 0 to 1, where 1 represents a perfect classifier and 0.5 represents a random classifier. As seen in (3.1) it's calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. TPR and FPR are defined as:

$$TPR = \frac{TP}{TP + FN} \tag{3.2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.3}$$

Where TPR in 3.2 stands for True Positive Rate. TP is True Positives, FN is False Negatives. FPR in 3.3 stands for False Positive rate. FP is False Positives, and TN refers to True Negatives.

In the multi-class classifications, the AUC is computed using a one-vs-all approach, where the AUC for each class is calculated and then averaged. For all model experiments the ROC AUC Score function has been used to compute AUC for each classifications, as used by MedMNIST.

Accuracy (ACC) on the other hand, measures the proportion of correct predictions out of all predictions made, this provides model performance in terms of how many accurate predictions the model performed overall. It is defined as:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

For multi-class tasks, accuracy is calculated by comparing the predicted labels to the ground truth for each instance. In binary classification problems, a decision threshold (commonly 0.5) is applied to convert predicted scores into class labels. Although both AUC and ACC provide insight to performance of a model, its important to note that AUC scores are more regarded when it comes to medical image classifications rather than ACC. This is because ACC provides an accuracy score, which could not always indicate the true performance of a model, specifically in highly imbalanced datasets, such as medical image datasets, where the classification task involves classes that are not equally distributed.

3.5.2 Evaluation Methodology

MedMNIST's Evaluator class was utilized in this project to compute AUC and accuracy for each model, ensuring consistent evaluation practices across different experiments. The Evaluator requires ground truth labels and predicted scores to calculate these metrics. For each evaluation, the model's output logits were converted to probabilities using softmax (for multi-class) or sigmoid (for binary-class) functions. The predicted probabilities (y_score) were then passed to the Evaluator, along with the ground truth labels (y_true), to compute AUC and ACC:

- AUC: For binary-class tasks, the model's output probabilities for the positive class were extracted, and the AUC was computed using roc_auc_score. For multi-class tasks, a one-vs-all approach was employed, calculating the AUC for each class and averaging to get the final AUC score.
- ACC: The getACC function compared the predicted class labels (obtained by applying a decision threshold to the probabilities) with the ground truth labels to compute accuracy.

This evaluation process was applied consistently across all models, including ResNet, ViT-based architectures, MedViT, and fine-tuned MedCLIP models. However, for the Zero-Shot Prompt-Based Classification with MedCLIP, only accuracy was calculated, as AUC requires a continuous prediction score distribution derived from model training. For zero-shot evaluations, the MedCLIP model made predictions based solely on the similarity between image embeddings and textual prompts, and accuracy was calculated as the proportion of correct predictions over the total test set.

3.5.2.1 Integration with Weights & Biases

To track, visualize, and compare the performance of the models, Weights & Biases (WandB) were integrated into the evaluation pipeline. During each epoch, training and validation metrics such as loss, accuracy, and AUC were logged to WandB, for real-time monitoring of model performance. For each evaluation phase (validation and test), AUC and accuracy scores were recorded. This allowed for a comprehensive comparison across different models and configurations.

3.5.2.2 Evaluation Strategy

The evaluation process was carefully structured to ensure consistent assessment of each model's performance, while also aligning with MedMNIST's evaluation strategy. During the training phase, AUC and accuracy were computed at the end of each epoch using the validation set. This was done with the Evaluator class, this was useful to monitor the model's generalization capability and adjust hyperparameters as needed. Integration with WandB enabled logging of these metrics, providing visualizations for the training process and allowing for detailed analysis afterwards.

Once training was completed, the best-performing model, as determined by its validation AUC, was evaluated on the test set to obtain the final AUC and accuracy scores. The Evaluator class was used for this purpose, ensuring that the results were directly comparable with MedMNIST's benchmark evaluations.

For the zero-shot evaluation with MedCLIP, I began by testing individual images, displaying both the predicted and actual labels to verify the model's predictions visually. Following this, I iterated over the entire test set to calculate accuracy based on the proportion of correct predictions. Additionally, I explored various text prompts to examine their influence on classification results. This method allowed me to assess how well MedCLIP could generalize to new image data without additional training, offering valuable insights into the model's versatility in a zero-shot setting.

3.6 Summary

This chapter outlined the methodology used to evaluate and extend medical image classification models on the MedMNISTv2 dataset collection. It begins with a detailed overview of the datasets, including PathMNIST, PneumoniaMNIST, and VesselMNIST3D, selected to test model performance across varied modalities and tasks. Standardized preprocessing, as set by MedMNIST, was maintained to ensure fair comparisons, with slight adjustments made for specific model requirements.

The chapter then describes the implementation of various model architectures, starting with the reproduction of MedMNIST benchmarks using ResNet18, ResNet50 and Resnet18 + ACS convolutions for VesselMNIST3D. Advanced models, including Vision Transformers (ViT) for both 2D and 3D data, MedViT, and MedCLIP, were explored. A joint 3D model combining ResNet18 with ACS convolutions and a custom 3D ViT was also developed to capture diverse spatial features.

Training procedures, hyperparameter choices, and data handling techniques are discussed in detail. Evaluation was conducted using AUC and accuracy metrics, ensuring alignment with MedMNIST's standards. Additionally, zero-shot prompt-based classification with MedCLIP was explored, offering insights into the model's potential to generalize without further training.

4 EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the results of various experiments conducted on each selected MedMNISTv2 and MedMNIST+ dataset: PneumoniaMNIST, PathMNIST, and VesselMNIST3D. For each dataset, the experimental results cover the reproduction of benchmarks, implementation of new model architectures (such as ViT and MedCLIP), and evaluations using different hyperparameters, including prompt-based zero-shot classification for the vision-language foundation model.

4.1 Experimental Results

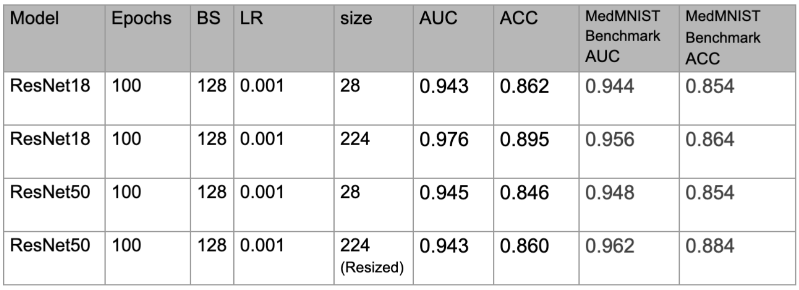
4.1.1 Benchmark Reproductions

4.1.1.1 PneumoniaMNIST

The reproduction of the benchmark results involved training and testing ResNet-18 and ResNet-50 on PneumoniaMNIST. The benchmark in the MedMNIST paper was achieved with Google AutoML, but here, ResNet models were employed to reproduce the results. ResNet-18 was trained using both image sizes 28 and 224, while ResNet-50 was trained with image size 28 and then resized to 224 due to computational limitations.

Table 4.1: Benchmark Performance achieved by MedMNIST for PneumoniaMNIST. BS: Batch Size, LR: Learning rate

Model	Epochs	BS	LR	Size	AUC	ACC
Google AutoML	100	128	0.001	28	0.991	0.946



Model	Epochs	BS	LR	size	AUC	ACC	MedMNIST Benchmark AUC	MedMNIST Benchmark ACC
ResNet18	100	128	0.001	28	0.943	0.862	0.944	0.854
ResNet18	100	128	0.001	224	0.976	0.895	0.956	0.864
ResNet50	100	128	0.001	28	0.945	0.846	0.948	0.854
ResNet50	100	128	0.001	224 (Resized)	0.943	0.860	0.962	0.884

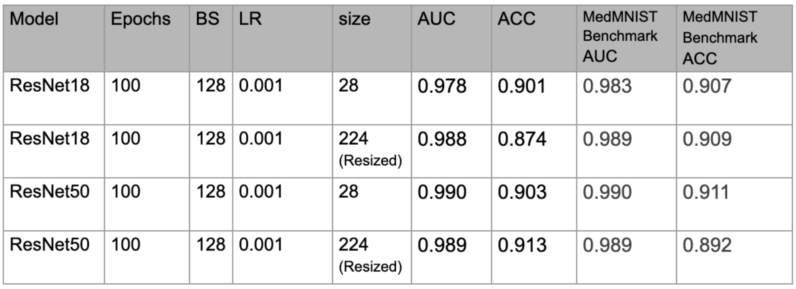
Figure 4.1: Reproducing MedMNIST Benchmark Models for PneumoniaMNIST

4.1.1.2 PathMNIST

The benchmark for PathMNIST was reproduced using ResNet18 and ResNet50 on image size 28. Due to the computational intensity of PathMNIST with the benchmarked models, reproducing benchmark results using image size 224 were outside of the available resources, therefore, image size 28 were resized using MedMNIST's recommended resize function to produce 224 size images and were then used to reproduce the benchmark. The models were trained successfully, and the results were consistent with the benchmarks reported in the MedMNIST paper.

Table 4.2: Benchmark Performance achieved by MedMNIST for PathMNIST. BS: Batch Size, LR: Learning rate

Model	Epochs	BS	LR	Size	AUC	ACC
ResNet50 (28)	100	128	0.001	28	0.990	0.911

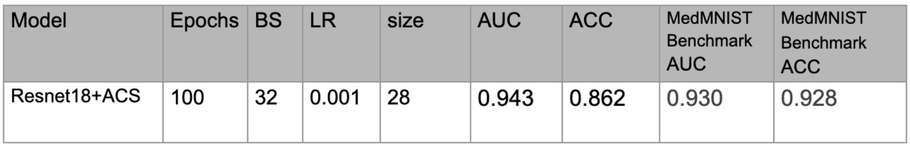


Model	Epochs	BS	LR	size	AUC	ACC	MedMNIST Benchmark AUC	MedMNIST Benchmark ACC
ResNet18	100	128	0.001	28	0.978	0.901	0.983	0.907
ResNet18	100	128	0.001	224 (Resized)	0.988	0.874	0.989	0.909
ResNet50	100	128	0.001	28	0.990	0.903	0.990	0.911
ResNet50	100	128	0.001	224 (Resized)	0.989	0.913	0.989	0.892

Figure 4.2: Reproducing MedMNIST Benchmark Models for PathMNIST

4.1.1.3 VesselMNIST3D

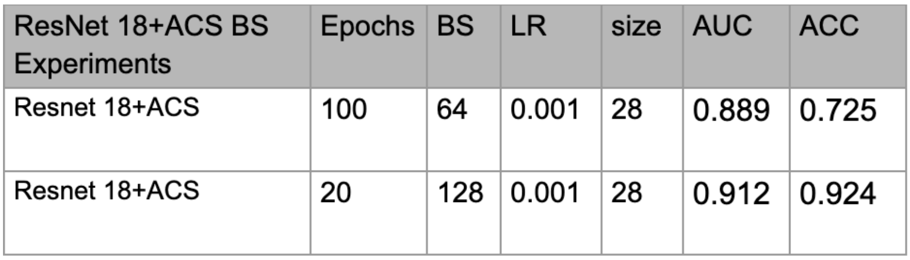
The benchmark was reproduced successfully using ResNet-18 + ACS convolutions.



Model	Epochs	BS	LR	size	AUC	ACC	MedMNIST Benchmark AUC	MedMNIST Benchmark ACC
Resnet18+ACS	100	32	0.001	28	0.943	0.862	0.930	0.928

Figure 4.3: Reproducing MedMNIST Benchmark Result achieved on VesselMNIST3D

Further experimentation on different batch sizes was done to observe their effects on the 3D model's performance.



ResNet 18+ACS BS Experiments	Epochs	BS	LR	size	AUC	ACC
Resnet 18+ACS	100	64	0.001	28	0.889	0.725
Resnet 18+ACS	20	128	0.001	28	0.912	0.924

Figure 4.4: Different Batch Size experimentation using Benchmark Model on VesselMNIST3D

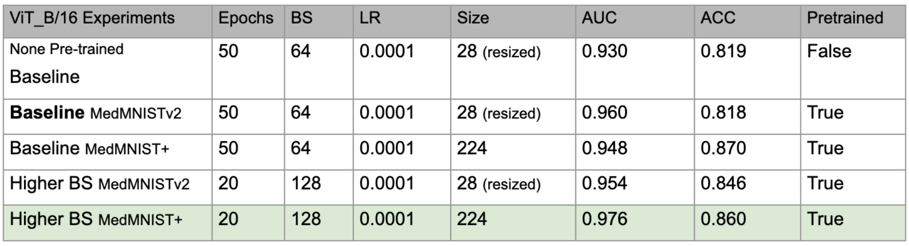
4.1.2 Vision Transformer Experiments

4.1.2.1 PneumoniaMNIST

The ViT model (vit_B/16 from timm) was implemented with various experiments:

- Non-Pretrained ViT: A non-pretrained version of ViT was trained with the baseline hyperparameters crafted for ViT experiments outlined in methodology section.
- Baseline: Pre-trained ViT was then fine-tuned on both image sizes, size 28 from MedMNISTv2 and size 224 from MedMNIST+ using the crafted baseline hyperparameters defined in the methodology section.
- Batch Size Experiments: In the next experiment, the batch size was increased to 128, and the model was fine-tuned again on both image sizes. This adjustment was done to assess the impact of batch size on model performance.

The hyperparameters used initially were as defined in the methodology section. However, due to higher computational complexity, the number of epochs was decreased to 20 when experimenting with higher batch sizes. The results of the experiments can be found in the table figure 4.5 below.



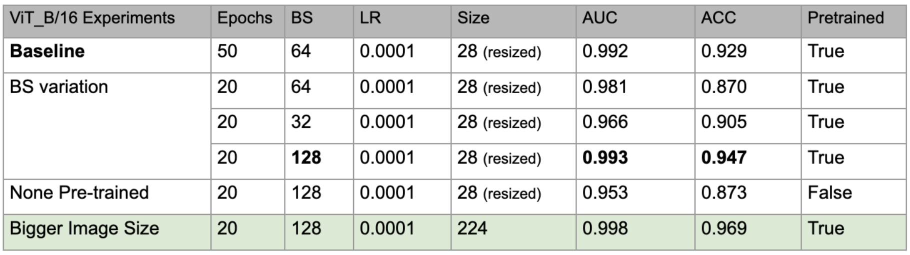
ViT_B/16 Experiments	Epochs	BS	LR	Size	AUC	ACC	Pretrained
None Pre-trained Baseline	50	64	0.0001	28 (resized)	0.930	0.819	False
Baseline MedMNISTv2	50	64	0.0001	28 (resized)	0.960	0.818	True
Baseline MedMNIST+	50	64	0.0001	224	0.948	0.870	True
Higher BS MedMNISTv2	20	128	0.0001	28 (resized)	0.954	0.846	True
Higher BS MedMNIST+	20	128	0.0001	224	0.976	0.860	True

Figure 4.5: ViT Fine-Tuning Experiments for PneumoniaMNIST

4.1.2.2 PathMNIST

Similar to PneumoniaMNIST, the ViT model was experimented with in a few ways for PathMNIST. The baseline model was trained for 50 epochs, but due to heavy computational complexity and lack of resources, further experiments were trained on 20 epochs.

- Baseline: ViT was trained using the methodology-defined hyperparameters on image size 28.
- Batch Size Variation: A series of experiments were conducted with batch sizes of 32, 64, and 128. The best result was obtained using a batch size of 128, therefore batch size 128 was used to compare performance with other experiments within the ViT experiments.
- Non-Pretrained ViT: To compare the effect of pre-training, the non-pretrained ViT was ran using the same hyperparameters as the best experiment above, the higher batch size.
- Limited Image Size 224 testing: Due to PathMNIST's large size, only the best experiment was compared to the trained image size 224 with the same hyperparameters.



ViT_B/16 Experiments	Epochs	BS	LR	Size	AUC	ACC	Pretrained
Baseline	50	64	0.0001	28 (resized)	0.992	0.929	True
BS variation	20	64	0.0001	28 (resized)	0.981	0.870	True
	20	32	0.0001	28 (resized)	0.966	0.905	True
	20	**128**	0.0001	28 (resized)	**0.993**	**0.947**	True
None Pre-trained	20	128	0.0001	28 (resized)	0.953	0.873	False
Bigger Image Size	20	128	0.0001	224	0.998	0.969	True

Figure 4.6: ViT Fine-Tuning Experiments for PathMNIST

4.1.2.3 VesselMNIST3D

For the VesselMNIST3D dataset, a custom ViT3D model was implemented using the baseline hyperparameters detailed in the methodology section. However, the initial experiments revealed a significant drop in the performance as compared to the benchmarked Resnet18 model, this was due to the significant class imbalance in VesselMNIST3D. To address this, nine additional experiments were conducted, each exploring different techniques for class imbalance mitigation. These techniques included oversampling the minority classes, applying various data augmentation strategies, and adjusting the loss function to incorporate class weights. The goal was to enhance the model's ability to correctly classify instances from the underrepresented classes, thereby improving its overall robustness and accuracy.

The first strategy for addressing the poor performance caused by class imbalance involved oversampling the minority class, which involved manually duplicating instances of the minority class within the training data using a custom dataset class. This oversampling method was the only approach that showed a noticeable improvement in the model's performance, mitigating some of the effects of class imbalance that were amplified when the ViT architecture was implemented.

In addition to oversampling, a series of augmentations using the MONAI library were applied to the dataset [8]. These augmentations included random flipping, affine transformations, elastic deformations, Gaussian noise addition, and intensity normalization, both collectively and in isolation. Visualizations of the images before and after augmentation (which are slices extracted from the original 3D medical images) were created to understand their impact on the dataset's characteristics. However, these augmentations did not result in performance gains. Similarly, implementing a weighted loss function to penalize misclassification of the minority class also failed to improve accuracy. Combining oversampling with MONAI augmentations or weighted loss only led to marginal improvements, further underscoring the difficulty of addressing class imbalance with these techniques. Images of each class before and after MONAI augmentations can be seen below.

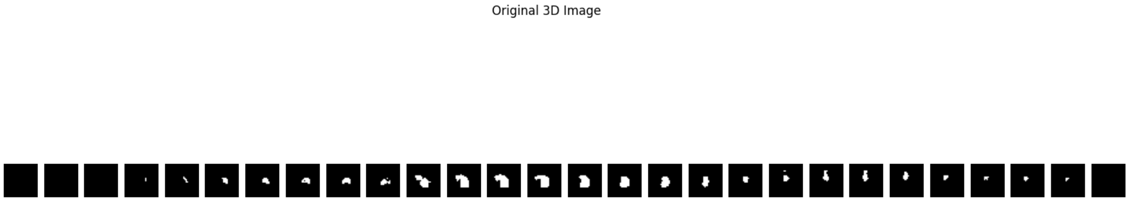


Figure 4.7: Slices from a Vessel Image.

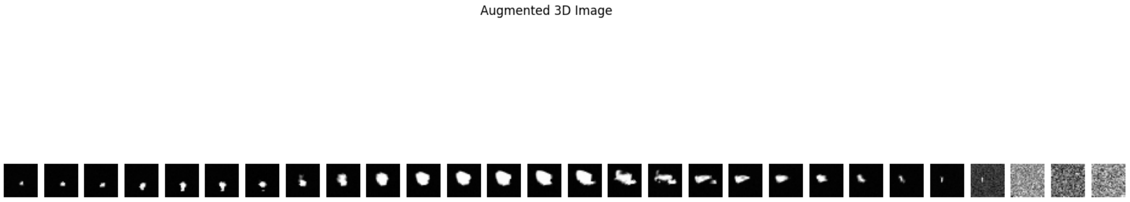


Figure 4.8: Slices from a Vessel Image after MONAI Augmentations.

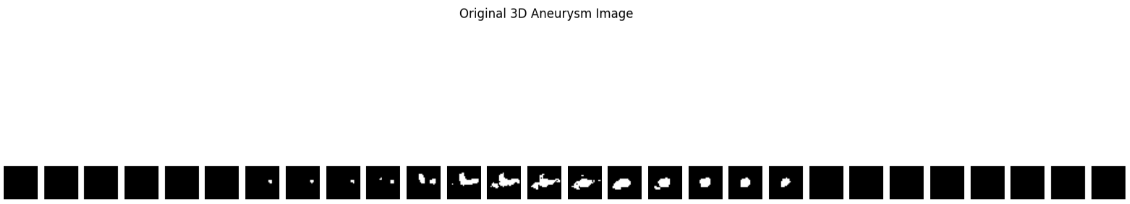


Figure 4.9: Slices from an Aneurysm Image.

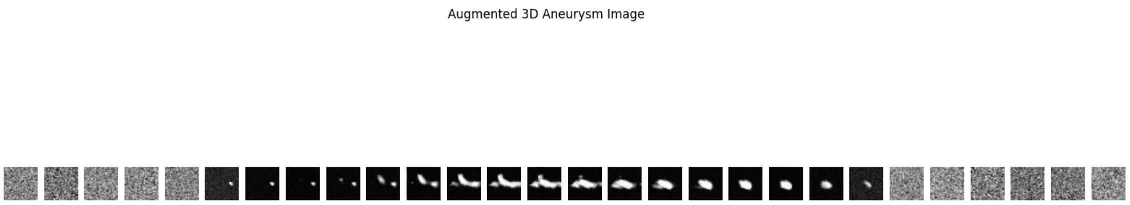
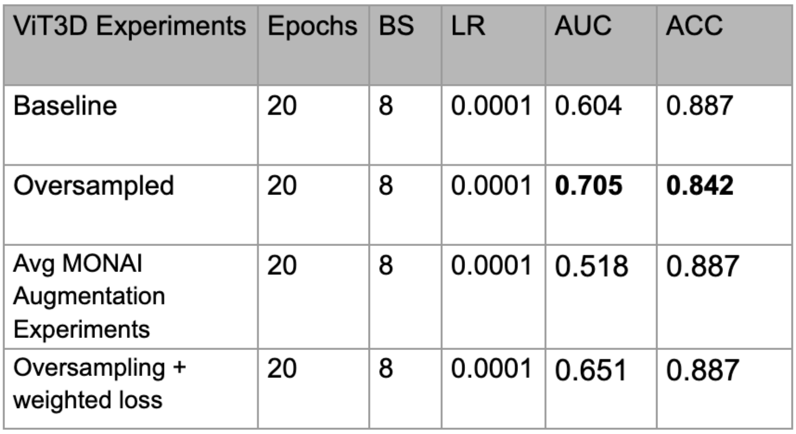


Figure 4.10: Slices from an Aneurysm Image after MONAI Augmentations.

One promising result emerged from a final experiment that involved combining oversampling with a custom loading of pre-trained weights from a 2D ViT model into the 3D ViT model. By rearranging the code to incorporate the 2D ViT's weights, the 3D ViT effectively became "pre-trained," which led to better performance than any of the previous approaches.



ViT3D Experiments	Epochs	BS	LR	AUC	ACC
Baseline	20	8	0.0001	0.604	0.887
Oversampled	20	8	0.0001	**0.705**	**0.842**
Avg MONAI Augmentation Experiments	20	8	0.0001	0.518	0.887
Oversampling + weighted loss	20	8	0.0001	0.651	0.887

Figure 4.11: Performance Result from ViT3D experiments on VesselMNIST3D

4.1.3 MedViT

4.1.3.1 PneumoniaMNIST

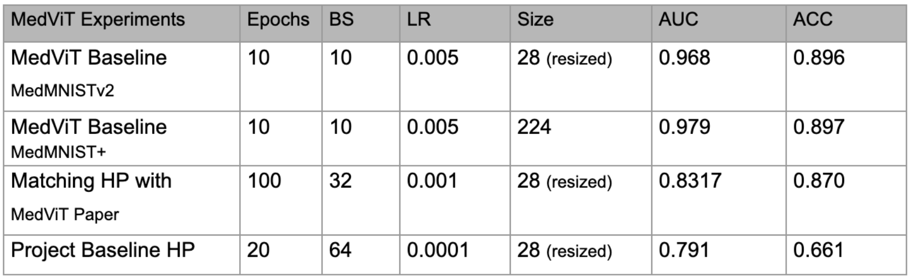
Multiple experiments were conducted with MedViT:

- MedViT Baseline: Initially, MedViT was fine-tuned with the hyperparameters recommended in its original implementation guide for both image sizes. The hyperparameters recommended to start with for fine-tuning can be found in the table below:

Table 4.3: Initial Hyperparameter Settings for fine-tuning MedViT . BS: Batch Size, LR: Learning rate

Epochs	LR	BS	Weight Decay	Optimizer
10	0.005	10	1e-2	SGD

- Hyperparameter Matching: The model was then trained using hyperparameters that matched those used in the MedMNIST benchmarks.
- Custom ViT Hyperparameters: Finally, MedViT was trained using the custom hyperparameters defined for ViT in the methodology section.



MedViT Experiments	Epochs	BS	LR	Size	AUC	ACC
MedViT Baseline MedMNISTv2	10	10	0.005	28 (resized)	0.968	0.896
MedViT Baseline MedMNIST+	10	10	0.005	224	0.979	0.897
Matching HP with MedViT Paper	100	32	0.001	28 (resized)	0.8317	0.870
Project Baseline HP	20	64	0.0001	28 (resized)	0.791	0.661

Figure 4.12: MedViT experiments for PneumoniaMNIST. BS: Batch Size, LR: Learning rate, HP: Hyperparameters

4.1.3.2 PathMNIST

MedViT was fine-tuned on PathMNIST using the smaller hyperparameters recommended in the MedViT implementation, shown in table 4.3, focusing on image size 28 due to dataset constraints. Due to the long training time, and lack of resources, for PathMNIST, only the smaller, recommended starting settings where implemented, as well as the Hyperparameter settings recommended by the MedViT paper which matches MedMNIST hyperparameters.

Table 4.4: Performance result for Fine-tuning MedViT on PathMNIST

MedViT Experiment	Epochs	BS	LR	Size	AUC	ACC
MedViT Baseline	10	10	0.005	28(resized)	**0.992**	**0.917**
Matching HP with MedViT paper	100	32	0.001	28(resized)	0.831	0.789

4.1.4 MedCLIP

4.1.4.1 PneumoniaMNIST

MedCLIP was fine-tuned using baseline hyperparameters for both image sizes of 28 resized to 224 and image size 224.

Table 4.5: MedCLIP Fine-Tuning Performance on PneumoniaMNIST. BS: Batch Size, LR: Learning rate

Model	Epochs	BS	LR	Size	AUC	ACC
MedCLIP	20	64	0.0001	28	0.986	0.921
MedCLIP	50	64	0.0001	224	**0.992**	**0.933**

4.1.4.2 PathMNIST

MedCLIP was fine-tuned for PathMNIST, and experimented in two ways:

1. Baseline: MedCLIP was fine-tuned using the baseline hyperparameters on image size 28 resized to 224 using MedMNIST's resize technique.
2. Batch Size Increase: A higher batch size was then used, inspired by the improved performance resulting from increased batch size throughout ViT-based model experiments on PathMNIST, This resulted in the best performance observed, surpassing the benchmark.

Table 4.6: MedCLIP Fine-Tuning Performance on PathMNIST. BS: Batch Size, LR: Learning rate

Model	Epochs	BS	LR	Size	AUC	ACC
MedCLIP	50	64	0.0001	28	0.985	0.895
MedCLIP	20	128	0.0001	28	**0.999**	**0.964**

4.1.5 MedCLIP Zero-Shot Prompt-Based Classification

4.1.5.1 PneumoniaMNIST

Zero-shot classification was performed using MedCLIP on the entire test set, with multiple prompt variations:

- Simple Class Prompts: Basic prompts structured as a question that directly mentioned class names, e.g., "Is there pneumonia in this image?"
- Descriptive Prompts: More descriptive prompts, such as "This X-ray shows signs of pneumonia."
- CLIP-Style Prompts: Prompts structured similarly to how CLIP defines them, using phrases like "a photo of chest X-ray that shows signs of a disease."
- ClIP-style Prompts with specific medical context phrases: CLIP-style prompts, using phrases that provide more specific relevant medical context like "a photo that is infectious."

Table 4.7: MedCLIP Zero-shot Prompt-based classification performance based on different prompt styles on PneumoniaMNIST

Prompt-Style	ACC
Simple Class Prompt	0.62
Descriptive Class Prompt	0.66
CLIP-style Prompt	0.64
CLIP-style Prompt + Medical Context	**0.85**

4.1.5.2 PathMNIST

Zero-shot prompt-based classifications were experimented with using MedCLIP, a vision-language foundation model pre-trained on medical data. The objective was to classify images from the PathMNIST dataset without explicit model fine-tuning for each specific class. Instead, MedCLIP utilized prompts describing the expected features of each class, aligning its visual and textual embeddings to generate predictions. The methodology and results are detailed below:

Class Descriptions as Prompts For the zero-shot classification, descriptive prompts were created for each of the nine classes in PathMNIST, focusing on specific cellular and tissue characteristics. This decision was made after analyzing the effect of using different prompt styles for classifying PneumoniaMNIST, as shown in table 4.7. For example, the prompt for "adipose" described "large, empty-looking cells with thin cytoplasmic borders and separated by fine connective tissue." Similarly, "colorectal adenocarcinoma epithelium" was characterized as having "irregularly arranged epithelial cells with prominent nuclei and disorganized growth patterns." These prompts provided MedCLIP with a textual anchor to associate the visual input with the correct class label. These textual prompts were gathered through definitions of each one of the classes from various medical sources and journals[CITE color Pathology terms]. The zero-shot classification was then carried out across the entire test set of PathMNIST, where the model predicted each image's label based on the provided prompts. The experiment revealed that MedCLIP struggled with certain classes, especially those with overlapping visual features such as distinguishing "smooth muscle" from "cancer-associated stroma". Overall accuracy was calculated, along with per-class accuracy, to gauge the model's performance.

Prompt Variation and Refinement

Multiple variations and styles of prompts were tested to explore how different wordings and specificity levels affect MedCLIP's classification accuracy: Initially, prompts directly describing the class features without mentioning the explicit class labels. For instance, "a photo of large, empty-looking cells" was used for "adipose.". Despite these variations, the model frequently misclassified visually similar tissues, particularly in cases where the differences were subtle, such as "mucus" versus "debris." Below are the best performing class prompts set for each class label:

- adipose: a photo of large, empty-looking cells with thin cytoplasmic borders and separated by fine connective tissue.

- background: a photo of a uniform, featureless area with no discernible cellular or tissue structures.
- debris: a photo of disorganized, granular material with scattered cellular fragments, indicative of cellular debris.
- lymphocytes: a photo of densely packed small, round cells with darkly staining nuclei and minimal cytoplasm.
- mucus: a photo of dense, pale-staining, and somewhat amorphous material filling spaces, characteristic of mucus accumulation within the tissue.
- smooth muscle: a photo of elongated, spindle-shaped cells arranged in parallel with smooth, eosinophilic cytoplasm.
- normal colon mucosa: a photo of well-organized glandular structures with regular, columnar epithelial cells lining the crypts, showing a healthy and intact tissue architecture with no signs of abnormal growth or disruption.
- cancer-associated stroma: a photo of disorganized, fibrous tissue with variable cellularity and irregularly spaced spindle-shaped cells, indicative of a reactive stromal response often associated with malignant tumors.
- colorectal adenocarcinoma epithelium: a photo of irregularly arranged epithelial cells with prominent nuclei, disorganized growth patterns, and signs of glandular formation.

Cross-Referencing with GPT-4

To further understand MedCLIP's decision-making process and misclassification patterns, a cross-referencing experiment was conducted using OpenAI's GPT-4 [35]. This involved presenting GPT-4 with the same images and prompts used for MedCLIP to assess if a language model with different architecture and training methods would make similar errors. The results showed that GPT-4 often made the same misclassifications as MedCLIP, particularly with "cancer-associated stroma" and "smooth muscle." During the cross-referencing, ChatGPT-4 was provided with a two-step process:

1. Initial Classification: GPT-4 received the image and was asked to identify the correct label from the list of 9 labels from PathMNIST dataset based on its visual features.

2. Prompt Refinement: It was then provided with detailed descriptions of each class to see if its classification changed. Surprisingly, even with the contextual knowledge, GPT-4, like MedCLIP, frequently maintained its initial misclassifications.

Table 4.8: Comparison of Zero-shot Prompt-based classification between MedCLIP and GPT-4 on PathMNIST experiments

# of tests	Ground Truth	Misclassified by MedCLIP as	Misclassified by GPT-4 as	# of correct MedCLIP	# of correct GPT-4
10	adipose	lymphocytes	lymphocytes	3	5
10	background	adipose	smooth muscle	4	6
10	debris	colorectal-adenocarcinoma-epithelium	mucus	5	6
10	lymphocytes	debris	debris	2	4
10	mucus	colorectal adenocarcinoma epithelium	debris	3	5
10	smooth muscle	cancer-associated stroma	cancer-associated stroma	6	7
10	normal colon mucosa	colorectal adenocarcinoma epithelium	cancer-associated stroma	5	6
10	cancer-associated stroma	smooth muscle	smooth muscle	5	5
10	colorectal adenocarcinoma epithelium	debris	cancer-associated stroma	8	7

4.1.6 ResNet18 + ViT Joined Architecture

A joint architecture combining ResNet18 with ACS convolutions and a custom 3D Vision Transformer (ViT3D) was implemented to address the performance disparity observed on the VesselMNIST3D dataset. The model utilized the same hyperparameters as those used for training the benchmark ResNet. In this setup, the outputs of the modified ResNet18 and ViT3D were concatenated to form a single feature vector for final classification. This joint approach was run five times, and the average result across these experiments showed a slight improvement over the original benchmark set by MedMNIST.

Table 4.9: Performance of Joint ResNet18 + ViT3D for VesselMNIST3D

Model	Epochs	BS	LR	Size	AUC	ACC
ResNet18(ACS)+ViT3D	100	32	0.001	28	**0.941**	**0.921**

4.2 Analysis

4.2.1 Model-Specific Analysis

4.2.1.1 Benchmark Reproduction

PneumoniaMNIST

The PneumoniaMNIST a 2D, Binary Classification task consisting of X-ray images, showed generally comparable and better performance across most models, due to its binary classification nature and the clear visual distinctions between healthy and pneumonia-affected lungs in X-ray images. Another contributing factor for PneumoniaMNIST's ease for model adaptation and generalization is due to its manageable and lightweight size.

The results for ResNet-18 and ResNet-50 on PneumoniaMNIST aligned well with the original benchmarks, despite using a different training approach than the Google AutoML method employed in the MedMNIST paper. This revealed that CNN-based architectures can still perform exceptionally in small-scale medical image classification tasks. Table 4.2 showcases the high AUC and ACC scored achieved with the ResNet models, specifically ResNet18, using image size 224 from the MedMNIST+ collection, this can be linked to ResNet's capability to capture low-level features in chest X-ray images, a task traditional CNNs are known to perform well on [14]. However, due to high computational requirements and complexity, ResNet50 could not be reproduced with the given hyperparameters on image size 224, to mitigate this, image size 28 from the MedMNISTv2 collection was resized to 224 and ResNet50 was fine-tuned. The computational complexity and resources intensity of deeper CNNs are one of the contributing reasons for the current research to adopt more lightweight and less resource intensive alternative models for biomedical image classifications.

PathMNIST

The PathMNIST dataset, consisting of multi-class pathology slide images, presented a more complex classification challenge due to the variety and subtle differences among tissue types, as well as large size of the dataset, being the largest dataset provided by MedMNIST. However, the ResNet-18 and ResNet-50 achieved results reproducing the benchamrking results. These models showed the capability of CNNs to extract features efficiently from high-resolution pathology images, achieving high AUC and ACC scores as shown in Table 4.2. These high-performing results confirm that CNNs have been successfully adapted to texture-rich and intricate images such as pathology slides [28].

VesselMNIST3D

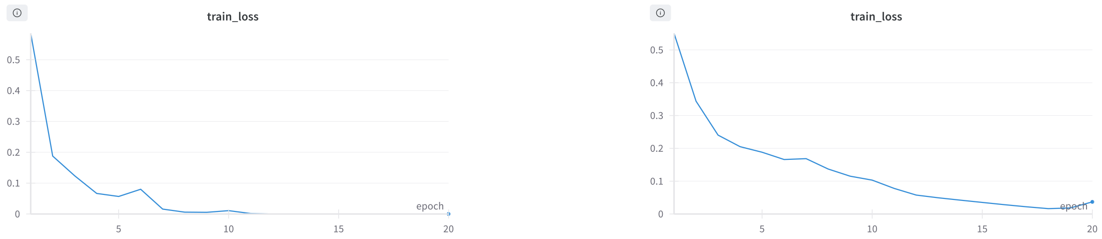
VesselMNIST3D had additional complexity with the 3D nature and class imbalance, consisting predominantly of vessel images with a minority class of aneurysms. ResNet-18 adapted with ACS convolutions was the highest-achieving benchmarked model for this dataset. Despite the class imbalance, ResNet-18 achieved a high AUC and ACC. Meaning that CNN-based architectures can be extended effectively into 3D medical image classification tasks [28]. This success is partly due to the ability of CNNs to extract 3D spatial features and utilize extensive contextual information, which is crucial in interpreting complex anatomical structures in brain MRA images.

The architecture's convolutional layers, designed to capture local patterns, helped in distinguishing between the vessel and aneurysm cases, even with the limited minority class. However, it is worth noting that the limitations of CNNs in handling severe imbalance are still present. This suggests that while CNNs offer a solid starting point, further adaptation or hybrid approaches, such as combining CNNs with transformer architectures, might be necessary to enhance performance in highly imbalanced 3D medical datasets. Further experiments with batch sizes were done with the benchmarked model to assess the effect of different batch sizes, as it can be seen in 4.4 showcasing that a batch size of 32 was an optimal batch size for the 3D dataset, as increasing batch size to 64 decreased performance, while increasing to 128 made performance better, while increasing resource requirements than of 64 but did not surpass the batch size of 32.

4.2.1.2 ViT Experiments

PneumoniaMNIST

The experiments with ViT revealed the importance of pretraining and batch size. A noticeable trend was the improvement in model performance with larger batch sizes, aligning with findings from Dosovitskiy et al. who noted that ViTs benefit from larger batch sizes due to their global attention mechanism [10]. The performance of ViT experiments can be seen in Table 4.5, where the performance gap between smaller and larger batch sizes can be seen through the AUC and ACC scores. Further, 4.13 showcases the difference in train loss between the two experiments. Where training loss with the highest performing ViT lessened as the epochs go on, while the non-pretrained ViT with the least AUC score, struggles with reducing the training loss, as it happens much slower over the epochs and as it reaches the last few epochs, it rises up again, suggesting that the model may have stopped learning.



(a) Train Loss for ViT with highest AUC

(b) Train Loss for non pre-trained ViT with lowest AUC

Figure 4.13: Comparison between training loss for the highest performing ViT and the lowest performing ViT for PneumoniaMNIST

PathMNIST

The ViT models showed increased performance on the PathMNIST dataset as compared to the highest benchmark, especially when using larger batch sizes and utilising pre-trained ViTs. This dataset's complexity, with its multiple classes and fine-grained differences in tissue structures, played well into the ViT's strengths. As outlined in Dosovitskiy et al., the global attention mechanism of ViTs enables the capture of long-range dependencies within images, making them particularly effective for complex medical images where spatial relationships are critical [10]. The experiments demonstrated that larger image sizes also improved the performance, which corresponds with findings that suggest ViTs benefit from high-resolution data, as they can process images in patch sequences, maintaining detailed spatial information. However, even though the highest performance achieved was with image size of 224, using image size of 28 and resizing to 224 as per ViT's input requirement still outperforms the benchmarked scores.

From the results shown in Table 4.6, it is clear that pretrained ViTs outperformed non-pretrained versions, further confirming the conclusion of Dosovitskiy et al. that large-scale pretraining can compensate for the lack of CNN-like inductive biases such as locality and translation equivariance in ViTs [10]. This is especially useful as ViTs suffer with a lack of inductive bias and using pretrained ViTs and fine-tuning on medical imaging tasks such as the multi-class classification with PathMNIST, allows the model to utilize the preexisting knowledge when it comes to low-level features, allowing it to learn better from the complex pathology slide images.

The validation loss curves for PathMNIST (Figure **??**) indicate that the pretrained ViT models achieve a more consistent reduction in loss over epochs, while non-pretrained models exhibited erratic loss patterns, showing difficulties in generalization. This behaviour further amplifies the importance of pretraining for ViTs, particularly in datasets requiring nuanced feature extraction, as highlighted in [27]. Additionally, the improved performance with larger batch sizes supports the conclusion that the global attention mechanism in ViTs can use more data during each training step, enhancing the model's capacity to generalize.

VesselMNIST3D

The baseline ViT3D model struggled to achieve comparable AUC scores, for which oversampling techniques helped. These results can be attributed to the fact that ViTs, ulike CNNs, which inherently capture local spatial information through convolutional filters, ViT3D relies on a global attention mechanism to process volumetric data, making it more susceptible to issues like class imbalance. The oversampling strategy improved model performance by enhancing exposure to minority class examples, allowing the ViT3D to better utilize its attention-based framework. However, considering VesselMNIST3D alone, ViT applications might not be the best fit for such a small and imbalanced dataset, unless it could be used in joint manner in hybrid models.

Despite the slight improvements with oversampling techniques, the ViT3D's performance on VesselMNIST3D did not match that of the hybrid ResNet18 + ViT3D model. This discrepancy can be linked to ViT's lack of inductive biases for vision tasks. While ViTs excel in large-scale pretraining, they do not inherently possess the locality and translation equivariance of CNNs [10]. Thus, incorporating CNN layers in a hybrid architecture helps in extracting local features more effectively from complex 3D medical images. This added benefit of the hybrid model can be seen in the Joint model architecture experiment performed for VesselMNIST3D in this project in 4.9, where AUC and ACC scores benefited the combination of CNN's local feature extraction with the ViT's global attention to handle the 3D medical data [27]. Further, standard data augmentations like rotation, flipping, and cropping did not significantly improve performance. Unlike CNNs, which benefit from localized changes, ViTs use a global attention mechanism that captures long-range dependencies. Therefore, augmentations that primarily introduce local changes does not necessarily alter the global context enough for ViTs to learn from effectively [10].

Additionally, in 3D medical imaging, augmentations that modify spatial configurations can distort anatomical structures, introducing noise instead of useful variability. This can be seen from the 4.8 and 4.10 where the slices from a vessel and an aneurysm have been shown before and after data augmentation. It can be seen that the data augmentation further distorts the slices and introduces noise to the image. Noting that transformers in medical imaging are sensitive to overall structure and context, so such disruptions can affect their ability to generalize [27].

MedViT: A Hybrid CNN-Transformer Architecture

Since MedViT was designed to bridge the gap between traditional CNNs and transformer-based models, its analysis and application for this project were crucial. MedViT combines the local feature extraction power of CNNs with the global attention mechanism of transformers [31]. The hybrid architecture integrates the strengths of both model types, allowing for robust performance in medical image classification tasks. In the experiments conducted, MedViT exhibited its best performance on the PneumoniaMNIST dataset, achieving an AUC of 0.979 and ACC of 0.897, which is slightly higher than the best results obtained with ViT, but not higher than MedMNIST's benchmark. PneumoniaMNIST's binary classification nature and relatively smaller size make it well-suited for MedViT's architecture. The incorporation of CNN-based locality in MedViT is an advantage for straightforward visual features of chest X-rays, such as edges and textures. This is aligned with the core design of MedViT, which includes convolutional blocks for efficient low-level feature extraction, while transformer blocks capture global relationships [31]. This blend of local and global feature processing contributes to MedViT's generalization capability in simpler tasks.

On the more complex PathMNIST dataset, which involves multiclass classification and contains diverse tissue structures, MedViT still performed well, achieving an AUC of 0.992 and an ACC of 0.917. However, it did not surpass the highest achievening ViT experiment. Interestingly, the experiments revealed that MedViT performed optimally with simpler hyperparameter settings, including smaller image sizes and batch sizes. This can be due to the inclusion of CNN components in its architecture. CNNs are inherently more efficient in capturing local textures and spatial hierarchies. [31]. Unlike ViTs, which requires larger batch sizes as seen from the experiments, MedViT's convolutional blocks allow for faster and more stable learning even with less complex hyperparameters. This property not only reduces computational demands but also enables the model to converge more rapidly, avoiding overfitting that can occur with excessive parameter tuning.

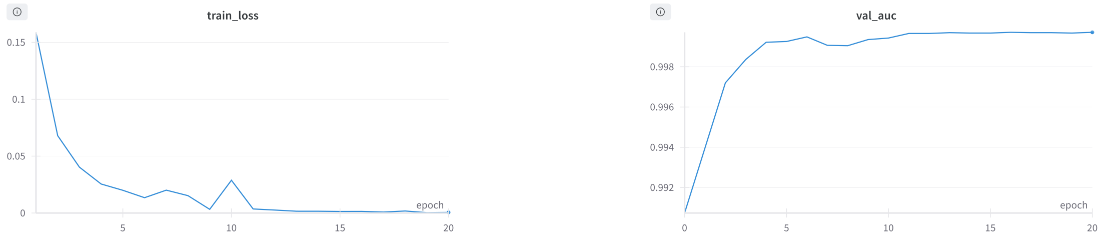
4.2.1.3 MedCLIP: A Fine-Tuned, Multimodal Approach

Fine-tuning MedCLIP led to exceptional results in the experiments, particularly on the PathMNIST dataset, where it achieved the highest AUC and ACC scores recorded for this study. This performance showcases the impact of MedCLIP's architecture, designed to integrate visual and textual medical information effectively, enhancing its representation capabilities and generalization [48].

During fine-tuning, MedCLIP's Vision Only model was used, because of the lack of textual dataset to go along the MedMNIST datasets. MedCLIP achieved high performance on both PneumoniaMNIST and PathMNIST, although its strengths were more pronounced with PathMNIST. On PneumoniaMNIST, MedCLIP reached an AUC of 0.992 and an ACC of 0.933, outperforming ViT and other models. This result could be due to MedCLIP's pretraining on large-scale medical image-text pairs, which provided the model with domain-specific knowledge of pneumonia-related features [48].

On the more complex PathMNIST dataset, MedCLIP surpassed all other models, getting the highest AUC of 0.999 and an ACC of 0.964. This improvement in performance can be linked to MedCLIP's multimodal pretraining, which involved decoupling images and texts, allowing the model to leverage a larger spectrum of medical data and semantic associations [48]. Unlike standard vision transformers or CNNs, MedCLIP's contrastive learning approach uses medical knowledge to align visual and textual representations. This alignment makes the model better understand the subtle differences between pathology images, effectively capturing the diverse tissue structures found in PathMNIST.

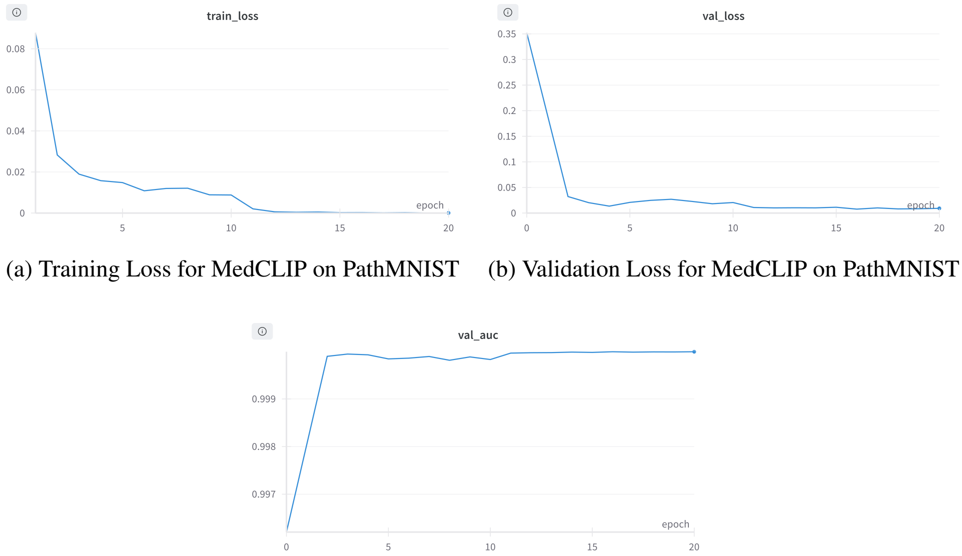
MedCLIP's performance with fine-tuning can also be explained by the model's use of "soft semantic matching loss," which helps in eliminating false negatives during contrastive learning [48]. This enables MedCLIP to maintain representation quality, thus explaining the improved performance across datasets. The training and validation loss curves for MedCLIP's fine-tuning on both datasets can be seen in 4.15 illustrating the smooth and rapid convergence, showcasing the efficacy of the multimodal learning framework.



(a) Training Loss for MedCLIP on PneumoniaMNIST

(b) Validation Loss for MedCLIP on PneumoniaMNIST

Figure 4.14: Training and Validation loss for MedCLIP on PneumoniaMNIST



(a) Training Loss for MedCLIP on PathMNIST

(b) Validation Loss for MedCLIP on PathMNIST

(c) Validation AUC for MedCLIP on PathMNIST

Figure 4.15: Training loss, Validation loss, and Validation AUC curves for best-performing experiments on PathMNIST

This performance supports the notion that multimodal models like MedCLIP, which incorporate both visual and textual information, offer significant advantages in medical image analysis. Fine-tuning allowed the model to adapt its rich, pre-trained embeddings to the specific dataset features, leading to groundbreaking results on PathMNIST and strong generalization on PneumoniaMNIST.

4.2.1.4 MedCLIP Zero-Shot Prompt-Based Classification

Zero-shot classification experiments with MedCLIP on both PneumoniaMNIST and PathMNIST revealed insights into how foundational vision-language models interpret medical images, especially in contexts where prior knowledge plays a significant role.

PneumoniaMNIST: Exploiting Medical Context

For PneumoniaMNIST, MedCLIP's zero-shot performance was impressive, particularly when the prompts used were constructed with a strong medical context. The model's success in this setting can be traced back to its pretraining on text generated from the CheXpert dataset, a rich source of medical reports that contain terms related to pneumonia and chest infections [48]. Since MedCLIP was trained to align visual and textual information specific to chest conditions, it recognized patterns in chest X-ray images when paired with relevant prompts. The model's training allowed it to understand nuanced clinical descriptions, making the accuracy score for the PneumoniaMNIST dataset high. the result of the performance can be seen in 4.7

MedCLIP's high performance is largely due to its pre-exposure to medical terminology related to during pretraining. When using more general or less descriptive prompts, its performance declined. This dependency on familiar medical contexts is in itself a limitations of zero-shot models in domains where the precise alignment of image features with domain-specific language is critical and sensitive. Therefore, MedCLIP's strong performance in the pneumonia classification task serves as an example showing the importance of domain-specific pretraining in enhancing zero-shot capabilities.

PathMNIST: The Struggle with Limited Knowledge

In contrast, MedCLIP struggled with the more complex, multiclass PathMNIST dataset during zero-shot classification. Despite the use of prompt engineering, MedCLIP was unable to accurately distinguish between the various tissue types present in colon pathology slides. This difficulty can be explained by the model's lack of prior exposure to pathology-specific knowledge during pretraining. Unlike its pretraining on chest X-ray reports, MedCLIP did not have access to the extensive and varied terminology needed to interpret the intricate visual patterns in pathology images [48]. As a result, the model showed poor generalization, often failing to capture the nuanced differences between similar classes.

Further complicating MedCLIP's performance, prior studies have highlighted specific misclassifications common in pathology images, particularly between muscle and stroma, and lymphocytes and debris/necrosis, when trained using CNN-based models [18]. This is mostly due to the fact that muscle and stroma share a fibrous architecture, therefore it's expected that their visual identifications require regor, which current vision-language foundation models do not possess on such sensitive and fine-grained medical image classification. This limitation points to the need for specialized pretraining in domains like pathology to better capture the subtleties inherent in such data.

Cross-Referencing with GPT-4: Insights into Model Mistakes

Cross-referencing the performance of MedCLIP with GPT-4 reveals additional insights into the nature of errors in zero-shot classification. According to the GPT-4 technical report, vision-language models like GPT-4 and MedCLIP are susceptible to "hallucinations" or inaccuracies when dealing with unfamiliar visual data [35]. In the case of PathMNIST, GPT-4 and MedCLIP struggled with similar misclassification patterns, indicating that these models may over-rely on surface-level features when they lack deeper domain-specific training. This observation tells us a key limitation of vision-language models: their performance is heavily influenced by the scope and quality of their pretraining data, as well as accurate and specific prompt-engineering for medical domain tasks.

The performance gaps seen in zero-shot tasks amplify that while models like MedCLIP can excel in contexts where they have been pretrained on relevant medical texts and images, they face challenges in domains requiring specialized knowledge, such as pathology. The result for the cross referencing of pathology classes can be found in 4.8

5 CONCLUSIONS

5.1 Conclusion

This dissertation aimed to extend the benchmarking landscape of the MedMNIST v2 dataset collection by evaluating the potential of various models, including vision transformers (ViTs), hybrid architectures, and vision-language foundation models to generalize across different medical imaging domains. By examining three distinct datasets: PneumoniaMNIST, PathMNIST, and VesselMNIST3D, this dissertation sought to address limitations in current benchmarks and explore alternative classification approaches. The results revealed both the strengths and limitations of these models, providing insights for future biomedical image analysis research.

The reproduction of benchmark models demonstrated that CNN-based architectures, particularly ResNet-18, remain effective for small-scale medical image classification tasks, especially for datasets like PneumoniaMNIST, where clear visual distinctions and manageable data sizes enable ease of adaptation and generalization. Despite their proven capabilities in extracting low-level features in medical images, CNNs exhibited challenges with computational complexity when scaling up to deeper architectures, such as ResNet-50, especially with larger image sizes. This limitation highlighted the need for lightweight models in resource-constrained environments.

ViT models offered a promising alternative, particularly in handling complex datasets such as PathMNIST. The experiments revealed the importance of large batch sizes and pretraining for achieving optimal performance, aligning with existing literature. However, ViTs faced difficulties with smaller, imbalanced 3D datasets, like VesselMNIST3D, suggesting that while transformers excel in capturing global spatial dependencies, they may struggle in scenarios where class imbalance and intricate 3D spatial structures are involved. This finding suggests the potential benefits of hybrid models that combine the strengths of CNNs and transformers.

MedViT, a hybrid CNN-transformer architecture, showcased robust performance in both binary and multi-class classification tasks, indicating its adaptability to varying dataset complexities. Its incorporation of convolutional layers aided in local feature extraction, while transformer blocks captured global relationships, balancing efficiency and generalization. MedViT's capacity to converge rapidly with simpler hyperparameters and smaller image sizes provides a practical advantage in biomedical image analysis, especially for smaller datasets.

MedCLIP, a vision-language foundation model pre-trained on medical data, demonstrated the highest performance on complex datasets like PathMNIST when fine-tuned, surpassing other models in terms of AUC and ACC scores. Its success highlighted the impact of multimodal pretraining. The zero-shot prompt-based experiments with MedCLIP further revealed its dependence on domain-specific pretraining, performing exceptionally on PneumoniaMNIST but struggling with the nuanced, multi-class nature of PathMNIST. This limitation pointed to the necessity for specialized pretraining in domains like pathology, where visual differences are subtle and contextually complex. Moreover, cross-referencing with GPT-4 confirmed the susceptibility of vision-language models to inaccuracies in unfamiliar contexts, reinforcing the critical role of comprehensive and targeted pretraining in medical imaging applications.

In conclusion, this dissertation provided a comprehensive evaluation of diverse model architectures for biomedical image classification, highlighting their varying strengths and limitations. While traditional CNNs remain reliable for simpler classification tasks, vision transformers and hybrid architectures like MedViT offer enhanced performance for complex datasets, dependent on adequate pretraining and careful selection of hyperparameters. Vision-language models like MedCLIP present new opportunities for multimodal learning but are heavily dependent on domain-specific knowledge and prompt engineering to achieve success. These findings advocate for a balanced approach in future research, integrating multiple model types and training strategies to address the diverse challenges inherent in biomedical image classification, and allows for further contributing and diversifying the benchmarking landscape for inclusive and lightweight datasets such as MedMNIST.

5.2 Evaluation

In this project, the objectives outlined in section 1.2 were successfully addressed, conducting evaluation of diverse deep learning models on three distinct MedMNIST v2 datasets. The experiments reproduced existing benchmarks using ResNet models, and extended the benchmarking work to newer models, such as Vision Transformers (ViTs) and MedViT. The ViT experiments demonstrated that with proper fine-tuning and adjustments in batch size and image dimensions, these models could surpass existing benchmarks, contributing valuable insights into their potential for medical image classification. Furthermore, the exploration of MedCLIP provided enhanced results, particularly on the PathMNIST dataset, highlighting the effectiveness of vision-language models in medical imaging. While the zero-shot prompt-based classification revealed exciting possibilities, I recognize that there was room for a deeper exploration of prompt design and more rigorous control over the evaluation environment. Similarly, the performance of ViT3D on the small, imbalanced VesselMNIST3D dataset indicated the need for further refinement in handling 3D data. Despite these limitations, I approached the project in a systematic and professional manner, ensuring alignment with MedMNIST's standard preprocessing steps for a fair comparison and meticulously evaluating the models' performances. Overall, this project has contributed to understanding how state-of-the-art models can diversify and enhance the medical image classification landscape, setting the stage for further exploration in this domain.

5.3 Limitations and Future Work

One limitation of this study is the scope of the Vision Transformer (ViT) experiments, which focused solely on the ViT_B/16 variant. While this provided valuable insights into the potential of ViTs for medical image classification, future work should explore other variants, such as ViTL/32 or hybrid transformer architectures that may better capture the complexities of different medical imaging modalities. The zero-shot prompt-based classification using MedCLIP was another area where limitations were identified. The cross-referencing conducted through GPT-4 used ChatGPT, which limited the control over the environment and analysis process. Future studies could utilize the GPT-4 API, allowing for a more precise and controlled exploration of prompt design and classification performance. Moreover, while MedCLIP excelled in vision-based classification tasks, its full potential was not realized due to the lack of accompanying textual reports in the MedMNIST datasets. A promising direction for future research would involve generating synthetic textual reports alongside image data, allowing for the use of MedCLIP in its intended multimodal form. This would enable a comprehensive evaluation of how integrating visual and textual data could further improve medical image classification. Additionally, exploring other foundation models beyond MedCLIP, including those pre-trained on a broader range of medical data, could reveal more about the adaptability and limitations of vision-language models in diverse medical domains. Lastly, the ViT3D experiments highlighted the challenges of working with small, imbalanced 3D datasets. Future work should focus on investigating more sophisticated approaches to address class imbalance to rigorously evaluate the capabilities of ViT architectures in handling 3D medical imaging tasks.

BIBLIOGRAPHY

[1] Alzubaidi, L. , Santamaría, J. , Manoufali, M. , Mohammed, B. , Fadhel, M. A. , Zhang, J. , Al-Timemy, A. H. , Al-Shamma, O. , and Duan, Y. . Mednet: Pre-trained convolutional neural network model for the medical imaging tasks, 2021. URL https://arxiv.org/abs/2110.06512.

[2] Bercovich, E. and Javitt, M. C. . Medical imaging: From Roentgen to the digital revolution, and beyond. *Rambam Maimonides Medical Journal*, 9(4):e0034, 10 2018. doi: 10.5041/rmmj.10355. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6186003/.

[3] Brown, T. B. , Mann, B. , Ryder, N. , Subbiah, M. , Kaplan, J. , Dhariwal, P. , Neelakantan, A. , Shyam, P. , Sastry, G. , Askell, A. , Agarwal, S. , Herbert-Voss, A. , Krueger, G. , Henighan, T. , Child, R. , Ramesh, A. , Ziegler, D. M. , Wu, J. , Winter, C. , Hesse, C. , Chen, M. , Sigler, E. , Litwin, M. , Gray, S. , Chess, B. , Clark, J. , Berner, C. , McCandlish, S. , Radford, A. , Sutskever, I. , and Amodei, D. . Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[4] Davenport, T. and Kalakota, R. . The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2):94–98, 6 2019. doi: 10.7861/futurehosp.6-2-94. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/.

[5] De Leucio, A. and De Jesus, O. . MR angiogram, 8 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK558984/.

[6] Devlin, J. , Chang, M.-W. , Lee, K. , and Toutanova, K. . Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[7] Djordjevic, I. B. . Chapter 14 - quantum machine learning. In Djordjevic, I. B. , editor, *Quantum Information Processing, Quantum Computing, and Quantum Error Correction (Second Edition)*, pages 619–701. Academic Press, second edition edition, 2021. ISBN 978-0-12-821982-9. doi: https://doi.org/10.1016/B978-0-12-821982-9.

00007-1. URL https://www.sciencedirect.com/science/article/pii/B9780128219829000071.

[8] Documentation, M. . Modules Overview — MONAI 0.8.0 Documentation, 2021. URL https://docs.monai.io/en/0.8.0/highlights.html#medical-image-data-i-o-processing-and-augmentation.

[9] Doerrich, S. , Salvo, F. D. , Brockmann, J. , and Ledig, C. . Rethinking model prototyping through the medmnist+ dataset collection, 2024. URL https://arxiv.org/abs/2404.15786.

[10] Dosovitskiy, A. , Beyer, L. , Kolesnikov, A. , Weissenborn, D. , Zhai, X. , Unterthiner, T. , Dehghani, M. , Minderer, M. , Heigold, G. , Gelly, S. , Uszkoreit, J. , and Houlsby, N. . An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[11] Goodfellow, I. , Bengio, Y. , and Courville, A. . *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[12] Goodfellow, I. J. , Pouget-Abadie, J. , Mirza, M. , Xu, B. , Warde-Farley, D. , Ozair, S. , Courville, A. , and Bengio, Y. . Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[13] Hajian-Tilaki, K. . Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation, 1 2013. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/.

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

[15] Hochreiter, S. . The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6:107–116, 04 1998. doi: 10.1142/S0218488598000094.

[16] Huix, J. P. , Ganeshan, A. R. , Haslum, J. F. , Söderberg, M. , Matsoukas, C. , and Smith, K. . Are natural domain foundation models useful for medical image classification?, 2023. URL https://arxiv.org/abs/2310.19522.

[17] Hussain, S. , Mubeen, I. , Ullah, N. , Shah, S. S. U. D. , Khan, B. A. , Zahoor, M. , Ullah, R. , Khan, F. A. , and Sultan, M. A. . Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical field: A review. *BioMed Research International*, 2022:1–19, 6 2022. doi: 10.1155/2022/5164970. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9192206/.

[18] Kather, J. N. , Krisam, J. , Charoentong, P. , Luedde, T. , Herpel, E. , Weis, C.-A. , Gaiser, T. , Marx, A. , Valous, N. A. , Ferber, D. , Jansen, L. , Reyes-Aldasoro, C. C. , Zörnig, I. , Jäger, D. , Brenner, H. , Chang-Claude, J. , Hoffmeister, M. , and Halama, N. . Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730, 1 2019. doi: 10.1371/journal.pmed.1002730. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6345440/.

[19] Kim, H. E. , Cosa-Linan, A. , Santhanam, N. , Jannesari, M. , Maros, M. E. , and Ganslandt, T. . Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1), 4 2022. doi: 10.1186/s12880-022-00793-7. URL https://doi.org/10.1186/s12880-022-00793-7.

[20] Kim, H. E. , Cosa-Linan, A. , Santhanam, N. , Jannesari, M. , Maros, M. E. , and Ganslandt, T. . Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1), 4 2022. doi: 10.1186/s12880-022-00793-7. URL https://doi.org/10.1186/s12880-022-00793-7.

[21] Kim, J. W. , Urooj Khan, A. , and Banerjee, I. . Systematic review of hybrid vision transformer architectures for radiological image analysis. *medRxiv*, 2024. doi: 10.1101/2024.06.21.24309265. URL https://www.medrxiv.org/content/early/2024/06/22/2024.06.21.24309265.

[22] Kohli, M. D. , Summers, R. M. , and Geis, J. R. . Medical image data and datasets in the era of machine learning—whitepaper from the 2016 c-mimi meeting dataset session. *Journal of Digital Imaging*, 30:392 – 399, 2017. URL https://api.semanticscholar.org/CorpusID:2290275.

[23] Krizhevsky, A. , Sutskever, I. , and Hinton, G. E. . Imagenet classification with deep convolutional neural networks. In Pereira, F. , Burges, C. , Bottou, L. , and Weinberger, K. , editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates,

Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[24] Lecun, Y. , Bottou, L. , Bengio, Y. , and Haffner, P. . Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[25] LeCun, Y. , Cortes, C. , and J.C. Burges, C. . MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges, 1998. URL https://yann.lecun.com/exdb/mnist/.

[26] LeCun, Y. , Bengio, Y. , and Hinton, G. . Deep learning. *Nature*, 521(7553):436–444, 5 2015. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539.

[27] Li, J. , Chen, J. , Tang, Y. , Wang, C. , Landman, B. A. , and Zhou, S. K. . Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives, 2022. URL https://arxiv.org/abs/2206.01136.

[28] Li, Q. , Cai, W. , Wang, X. , Zhou, Y. , Feng, D. D. , and Chen, M. . Medical image classification with convolutional neural network. In *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 844–848, 2014. doi: 10.1109/ICARCV.2014.7064414.

[29] Liu, X. , Faes, L. , Kale, A. U. , Wagner, S. K. , Fu, D. J. , Bruynseels, A. , Mahendiran, T. , Moraes, G. , Shamdas, M. , Kern, C. , Ledsam, J. R. , Schmid, M. K. , Balaskas, K. , Topol, E. J. , Bachmann, L. M. , Keane, P. A. , and Denniston, A. K. . A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019. ISSN 2589-7500. doi: https://doi.org/10.1016/S2589-7500(19)30123-2. URL https://www.sciencedirect.com/science/article/pii/S2589750019301232.

[30] Magee, D. , Treanor, D. , Crellin, D. , Shires, M. , Smith, K. , Mohee, K. , and Quirke, P. . Colour normalisation in digital histopathology images. *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, 01 2009.

[31] Manzari, O. N. , Ahmadabadi, H. , Kashiani, H. , Shokouhi, S. B. , and Ayatollahi, A. . Medvit: A robust vision transformer for generalized medical image classifica-

tion. *Computers in Biology and Medicine*, 157:106791, May 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.106791. URL http://dx.doi.org/10.1016/j.compbiomed.2023.106791.

[32] Milletari, F. , Navab, N. , and Ahmadi, S.-A. . V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016. URL https://arxiv.org/abs/1606.04797.

[33] Miotto, R. , Wang, F. , Wang, S. , Jiang, X. , and Dudley, J. T. . Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19 6:1236–1246, 2018. URL https://api.semanticscholar.org/CorpusID:2740197.

[34] Montalbo, F. J. P. , Hernandez, L. R. T. , Palad, L. P. , Castillo, R. C. , Alon, A. S. , and De Ocampo, A. L. P. . Performance analysis of lightweight vision transformers and deep convolutional neural networks in detecting brain tumors in mri scans: An empirical approach. In *Proceedings of the 2023 8th International Conference on Biomedical Imaging, Signal Processing*, ICBSP '23, page 17–25, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716584. doi: 10.1145/3634875.3634878. URL https://doi.org/10.1145/3634875.3634878.

[35] OpenAI, Achiam, J. , Adler, S. , Agarwal, S. , Ahmad, L. , Akkaya, I. , Aleman, F. L. , Almeida, D. , Altenschmidt, J. , Altman, S. , Anadkat, S. , Avila, R. , Babuschkin, I. , Balaji, S. , Balcom, V. , Baltescu, P. , Bao, H. , Bavarian, M. , Belgum, J. , Bello, I. , Berdine, J. , Bernadett-Shapiro, G. , Berner, C. , Bogdonoff, L. , Boiko, O. , Boyd, M. , Brakman, A.-L. , Brockman, G. , Brooks, T. , Brundage, M. , Button, K. , Cai, T. , Campbell, R. , Cann, A. , Carey, B. , Carlson, C. , Carmichael, R. , Chan, B. , Chang, C. , Chantzis, F. , Chen, D. , Chen, S. , Chen, R. , Chen, J. , Chen, M. , Chess, B. , Cho, C. , Chu, C. , Chung, H. W. , Cummings, D. , Currier, J. , Dai, Y. , Decareaux, C. , Degry, T. , Deutsch, N. , Deville, D. , Dhar, A. , Dohan, D. , Dowling, S. , Dunning, S. , Ecoffet, A. , Eleti, A. , Eloundou, T. , Farhi, D. , Fedus, L. , Felix, N. , Fishman, S. P. , Forte, J. , Fulford, I. , Gao, L. , Georges, E. , Gibson, C. , Goel, V. , Gogineni, T. , Goh, G. , Gontijo-Lopes, R. , Gordon, J. , Grafstein, M. , Gray, S. , Greene, R. , Gross, J. , Gu, S. S. , Guo, Y. , Hallacy, C. , Han, J. , Harris, J. , He, Y. , Heaton, M. , Heidecke, J. , Hesse, C. , Hickey, A. , Hickey, W. , Hoeschele, P. , Houghton, B. , Hsu, K. , Hu, S. , Hu, X. , Huizinga, J. , Jain, S. , Jain, S. , Jang, J. , Jiang, A. , Jiang, R. , Jin, H. , Jin, D. , Jomoto, S. , Jonn, B. , Jun, H. , Kaftan, T. , Kaiser, Łukasz,

Kamali, A. , Kanitscheider, I. , Keskar, N. S. , Khan, T. , Kilpatrick, L. , Kim, J. W. , Kim, C. , Kim, Y. , Kirchner, J. H. , Kiros, J. , Knight, M. , Kokotajlo, D. , Kondraciuk, Łukasz, Kondrich, A. , Konstantinidis, A. , Kosic, K. , Krueger, G. , Kuo, V. , Lampe, M. , Lan, I. , Lee, T. , Leike, J. , Leung, J. , Levy, D. , Li, C. M. , Lim, R. , Lin, M. , Lin, S. , Litwin, M. , Lopez, T. , Lowe, R. , Lue, P. , Makanju, A. , Malfacini, K. , Manning, S. , Markov, T. , Markovski, Y. , Martin, B. , Mayer, K. , Mayne, A. , McGrew, B. , McKinney, S. M. , McLeavey, C. , McMillan, P. , McNeil, J. , Medina, D. , Mehta, A. , Menick, J. , Metz, L. , Mishchenko, A. , Mishkin, P. , Monaco, V. , Morikawa, E. , Mossing, D. , Mu, T. , Murati, M. , Murk, O. , Mély, D. , Nair, A. , Nakano, R. , Nayak, R. , Neelakantan, A. , Ngo, R. , Noh, H. , Ouyang, L. , O'Keefe, C. , Pachocki, J. , Paino, A. , Palermo, J. , Pantuliano, A. , Parascandolo, G. , Parish, J. , Parparita, E. , Passos, A. , Pavlov, M. , Peng, A. , Perelman, A. , Avila Belbute Peres, F. , de, Petrov, M. , Oliveira Pinto, H. P. , de, Michael, Pokorny, Pokrass, M. , Pong, V. H. , Powell, T. , Power, A. , Power, B. , Proehl, E. , Puri, R. , Radford, A. , Rae, J. , Ramesh, A. , Raymond, C. , Real, F. , Rimbach, K. , Ross, C. , Rotsted, B. , Roussez, H. , Ryder, N. , Saltarelli, M. , Sanders, T. , Santurkar, S. , Sastry, G. , Schmidt, H. , Schnurr, D. , Schulman, J. , Selsam, D. , Sheppard, K. , Sherbakov, T. , Shieh, J. , Shoker, S. , Shyam, P. , Sidor, S. , Sigler, E. , Simens, M. , Sitkin, J. , Slama, K. , Sohl, I. , Sokolowsky, B. , Song, Y. , Staudacher, N. , Such, F. P. , Summers, N. , Sutskever, I. , Tang, J. , Tezak, N. , Thompson, M. B. , Tillet, P. , Tootoonchian, A. , Tseng, E. , Tuggle, P. , Turley, N. , Tworek, J. , Uribe, J. F. C. , Vallone, A. , Vijayvergiya, A. , Voss, C. , Wainwright, C. , Wang, J. J. , Wang, A. , Wang, B. , Ward, J. , Wei, J. , Weinmann, C. , Welihinda, A. , Welinder, P. , Weng, J. , Weng, L. , Wiethoff, M. , Willner, D. , Winter, C. , Wolrich, S. , Wong, H. , Workman, L. , Wu, S. , Wu, J. , Wu, M. , Xiao, K. , Xu, T. , Yoo, S. , Yu, K. , Yuan, Q. , Zaremba, W. , Zellers, R. , Zhang, C. , Zhang, M. , Zhao, S. , Zheng, T. , Zhuang, J. , Zhuk, W. , and Zoph, B. . Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[36] Radford, A. , Kim, J. W. , Hallacy, C. , Ramesh, A. , Goh, G. , Agarwal, S. , Sastry, G. , Askell, A. , Mishkin, P. , Clark, J. , Krueger, G. , and Sutskever, I. . Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

[37] Rajpurkar, P. , Irvin, J. , Zhu, K. , Yang, B. , Mehta, H. , Duan, T. , Ding, D. , Bagul, A. , Langlotz, C. , Shpanskaya, K. , Lungren, M. P. , and Ng, A. Y. . Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.

org/abs/1711.05225.

[38] Ronneberger, O. , Fischer, P. , and Brox, T. . U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

[39] Rumelhart, D. E. , Hinton, G. E. , and Williams, R. J. . Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 10 1986. doi: 10.1038/323533a0. URL https://www.nature.com/articles/323533a0.

[40] Sarraf, S. and Tofighi, G. . Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks, 2016. URL https://arxiv.org/abs/1603.08631.

[41] Sarraf, S. and Tofighi, G. . Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks, 2016. URL https://arxiv.org/abs/1603.08631.

[42] Scabini, L. , Sacilotti, A. , Zielinski, K. M. , Ribas, L. C. , Baets, B. D. , and Bruno, O. M. . A comparative survey of vision transformers for feature extraction in texture analysis, 2024. URL https://arxiv.org/abs/2406.06136.

[43] Shen, D. , Wu, G. , and Suk, H.-I. . Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 6 2017. doi: 10.1146/annurev-bioeng-071516-044442. URL https://doi.org/10.1146/annurev-bioeng-071516-044442.

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

[45] Steiner, A. , Kolesnikov, A. , Zhai, X. , Wightman, R. , Uszkoreit, J. , and Beyer, L. . How to train your vit? data, augmentation, and regularization in vision transformers, 2022. URL https://arxiv.org/abs/2106.10270.

[46] Varoquaux, G. and Cheplygina, V. . Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digital Medicine*, 5, 2022. URL https://api.semanticscholar.org/CorpusID:232269760.

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

[48] Wang, Z. , Wu, Z. , Agarwal, D. , and Sun, J. . Medclip: Contrastive learning from unpaired medical images and text, 2022. URL https://arxiv.org/abs/2210.10163.

[49] Wu, B. , Xu, C. , Dai, X. , Wan, A. , Zhang, P. , Yan, Z. , Tomizuka, M. , Gonzalez, J. , Keutzer, K. , and Vajda, P. . Visual transformers: Token-based image representation and processing for computer vision, 2020.

[50] Yang, J. , Huang, X. , He, Y. , Xu, J. , Yang, C. , Xu, G. , and Ni, B. . Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, 25(8): 3009–3018, Aug. 2021. ISSN 2168-2208. doi: 10.1109/jbhi.2021.3049452. URL http://dx.doi.org/10.1109/JBHI.2021.3049452.

[51] Yang, J. , Shi, R. , Wei, D. , Liu, Z. , Zhao, L. , Ke, B. , Pfister, H. , and Ni, B. . MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1), 1 2023. doi: 10.1038/s41597-022-01721-8. URL https://www.nature.com/articles/s41597-022-01721-8.

[52] Zhang, H. and Qie, Y. . Applying deep learning to medical imaging: A review. *Applied Sciences*, 13(18), 2023. ISSN 2076-3417. doi: 10.3390/app131810521. URL https://www.mdpi.com/2076-3417/13/18/10521.