National College of Ireland

An MLHOps-Driven Vision Transformer Approach for Pneumonia Classification in Chest X-Rays

MSc Research Project
Cloud Computing

Dharma Teja Venkatesh Nagothi
Student ID: X22173897

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Dharma Teja Venkatesh Nagothi
Student ID:	X22173897
Program:	MSc Cloud computing
Year:	2024
Module:	MSc Research Project
Supervisor:	POFESSOR Vikas Sahni
Submission Due Date:	25/04/2024
Project Title:	An MLHOps-Driven Vision Transformer Approach for Pneumonia Classification in Chest X-Rays
Word Count:	7516
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Venkatesh
Date:	25/04/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	Venkatesh
Date:	25/04/24
Penalty Applied (if applicable):	

An MLHOps-Driven Vision Transformer Approach for Pneumonia Classification in Chest X-Rays

Dharma Teja Venkatesh Nagothi
X22173897

Abstract

Diagnostic innovations are needed to enhance early detection and treatment of pneumonia, a global public health crisis. This work improves pneumonia detection from chest X-rays using Machine Learning Operations (MLOps) and Vision Transformers (ViT), a sophisticated deep learning (DL) model used in many computer vision applications. Pre-processed chest X-rays are given to a ViT model for feature extraction. The ViT encoder extracts hierarchical visual cues, whereas the classifier predicts pneumonia. The proposed method is tested using 5863 pneumonia-labeled NIH Chest X-ray pictures. The experiment compares the ViT model to a CNN classifier model for pneumonia classification based on accuracy, sensitivity, specificity, and AUC score criteria and it was found that ViT performed in terms of validation accuracy of 95.22% better than CNN accuracy of 94.84%. The ViT model execution was completed much faster than CNN execution (3018 seconds vs 7816 seconds). Based on these results, ViT was chosen to be implemented using MLOps practices for model training, evaluation, and deployment on Microsoft AzureML cloud. The suggested pneumonia detection on ML Health Operations (MLHOps) infrastructure using integrated ML pipelines allows rapid iteration and model optimization and ensures reproducibility for additional medical image analysis applications.

1 Introduction

Pneumonia is a severe lung infection caused by several viral diseases. Pneumonia is the leading cause of infectious disease-related death in Western countries. Early detection and treatment of pneumonia can help manage the condition effectively. Approximately 800,000 children under the age of five die from pneumonia annually, equating to over 2,200 deaths every day. Pneumonia is detected by chest X-ray scans. Diagnosing Pneumonia from chest X-ray pictures is difficult even for skilled radiologists. Diagnosing and managing pneumonia on chest X-ray images can be challenging because of its resemblance to other lung conditions. Computerized assistance technologies are necessary to aid radiologists in diagnosing Pneumonia from chest X-ray pictures. Automated medical image classification has considerably expanded to diagnose medical images into predetermined classes. Deep Learning (DL) has become a popular technology for medical picture categorization jobs. DL models outperformed traditional methods when analyzing chest X-ray pictures of patients with pneumonia (Ben Atitallah et al.; 2022; Iori et al.; 2022).

Recent studies have shown significant advancements in DL models for classifying medical images and detecting objects. Convolutional neural networks (CNNs) have shown remarkable performance in identifying pneumonia from chest X-rays (Rajpurkar et al.;

2017; Li et al.; 2020). Most studies have mostly concentrated on maximizing accuracy metrics during internal validation. The CNN model examines the relationship between adjacent pixels within a specific receptive area defined by the filter size. Therefore, it is challenging to establish connections with remote pixels. There has been a lack of thorough evaluation to determine if these models will effectively apply to diverse real-world clinical environments. The lack of model explainability remains a barrier to clinical adoption (Wiens et al.; 2019). Transformers, a self-attention-based architecture, offer a new approach to address the visual complexities in modern medical image processing.

1.1 Problem Statement

The advancements in pneumonia detection models to facilitate the modern medical care system have been stunted by the lack of standardized development, evaluation, and governance protocols tailored to healthcare machine learning (Sendak et al.; 2020). Embracing the growing conventions around ML governance will be vital for gaining practitioner acceptance and enabling seamless incorporation with healthcare workflows that include model development, bias mitigation, monitoring, and maintenance – collectively termed as MLOps. (Alsagheer et al.; 2023). There is a need for not just accurate models, but accountable and reliable AI systems engineered for patient benefit. This provides the motivation for exploring alternate architectures better suited for medical visual data. Vision transformers (ViTs) have emerged as a promising candidate, demonstrating strong performance on imaging tasks while offering inherent advantages for contextual modeling and explainability (Parvaiz et al.; 2023).

1.2 Motivation

There is a compelling need for automated and reliable pneumonia screening to improve outcomes and save lives. We postulate that the convergence of two emerging techniques - Vision Transformers (ViTs) and MLOps for Health (MLHOps) could overcome many barriers to real-world clinical deployment. On the model architecture front, ViTs are well-suited to handle the visual complexity and contextual nuances of medical images. Additionally, ViTs segment images into patches and tokens in a way that is closer to natural language, lending themselves better to sequence-based explanations about model predictions than CNN-based approaches (Han et al.; 2022). Complementing these predictive advantages, orchestrating model development, evaluation, and deployment per MLOps guidelines could address pressing gaps in transparency, fairness, and monitoring.

1.3 Research Question

Can a vision transformer (ViT) model implemented as an end-to-end MLHOps pipeline on Azure outperform CNN models in accurately classifying chest X-ray images for pneumonia detection?

1.4 Research Objective

This thesis proposes a ViT architecture for classifying pneumonia from chest X-rays, assessed through detailed empirical evaluation and compared with CNN. The best-performing model was selected and implemented using MLHOps practices for automated model training, evaluation, and deployment using AzureML pipelines.

1.5 Research Contributions

The salient contributions of this research include:

1. Attaining state-of-the-art classification performance on public chest X-ray datasets through an improved ViT based deep learning approach.

2. Conducting laborious bias assessments on training data and internally testing model fairness across key demographic attributes to mitigate ethical hazards.

3. Operationalizing the ML lifecycle through MLHOps - leveraging data versioning, automated retraining pipelines, model deployment and continuous monitoring to ensure reliability in medical image analysis.

4. Providing a comprehensive evaluation that demonstrates the superior performance of ViTs over traditional CNN models.

1.6 Scope and Limitations

This research focuses on developing an end-to-end pipeline for automated pneumonia screening from chest radiographs. The findings may subsequently be extended to multi-disease classification, optimized deployment across varied healthcare environments, and combinations with clinical metadata analytics. However, such directions are beyond the current scope of this research undertaking. The model development and evaluation are limited to openly available chest X-ray datasets from journal publications and public repositories.

1.7 Thesis Structure

Section 1 presents the research projects, motivation, research question and objectives of the research. Section 2 details all the existing literature and their research gap that this research tries to cover. Section 3 provides the methodology for pneumonia detection using CNN and ViT models. Section 4 details the design specification of the ViT and CNN model architecture. Section 5 provides the implementation of CNN and ViT with training parameters and presents the MLHOps approach. Section 6 showcases the evaluation of both ViT and CNN in terms of accuracy, loss and other metrics. Section 7 concludes the research with future scope.

2 Related Work

Transformers have demonstrated similar performance to CNNs in tasks such as medical image classification (Matsoukas et al.; 2021), segmentation (Heidari et al.; 2023), and reconstruction (Zhou et al.; 2023). The present literature documents transformers with superior performance compared to even the most advanced CNN models. However, there is an ongoing debate over the performance of transformers against CNNs, with new improvements to transformer architecture introduced periodically to address transformer-related limitations. Dosovitskiy et al. (2020) introduced Vision Transformer (ViT), an adaptation of the original transformer model to patches of images in sequences to extract the salient information. It is made possible by the optimization of the attention

mechanism that combines the global context into visual features without compromising on computational efficiency. At present, researchers are further probing at the flexibility of vision transformers in addressing different problems in diverse fields. This section aims to offer a nuanced review of how ViTs contribute to the classification problem by automatically diagnosing diseases, specifically focusing on their efficacy in detecting and classifying pneumonia.

2.1 Medical Image Analysis using Deep Learning

For medical image segmentation, the authors (Jafari et al.; 2020) proposed DRU-Net, a novel deep convolutional neural network (DCNN) architecture that combines ResNet and DensNet's benefits. The skip connections and fewer model parameters used in comparison to DenseNet and attention networks (AttnNet) improved the segmentation accuracy with efficient training. The suggested method was tested on the Grand-challenge dataset for skin lesion segmentation and a local brain MRI dataset for multi-class segmentation. DRU-Net also outperformed ResNet, DenseNet, and AttnNet in segmentation accuracy, precision, and Jaccard metrics eboth the datasets, especially for label classes with few pixels and training examples. A similar research proposed in (Feng et al.; 2020) using DCNN with a cascading structure and conditional random fields (CRF) was used to segment medical images. The CRF is used for post-segmentation processing to resolve the conflict between segmentation accuracy and network depth. The CNN-CRF model's cascading structure simulates spatial closure tag dependencies and enhances segmentation accuracy. The research presented in (An et al.; 2021) used multiscale CNN (MCNN) with visual attention for medical image classification. The technique extracts high-level discriminative features and uses a Mahalanobis distance optimization model for robust training to enhance image classification. The algorithm was evaluated on the JSRT lung nodule database and the WBCD breast cancer database. It outclassed traditional DL methods in classification accuracy owing to the visual attention mechanism combined with unique MCNN architecture increasing the performance, proving the algorithm's stability and robustness in medical image classification.

Tang et al. (2022) introduced a self-supervised learning system for 3D medical image analysis using Swin UNETR, a unique 3D transformer-based model. To improve 3D CT image analysis, the model was pre-trained using 5,050 publically available CT scans from diverse body organs. The pre-trained Swin UNETR model performed well in segmentation tasks involving 13 abdominal organs, proving that self-supervised pre-training improves downstream segmentation tasks and opens up new uses for large-scale unlabeled medical image datasets. In (Zeid et al.; 2021), Vision Transformers (ViT) were used to classify colorectal cancer (CRC) histology images, proving their efficacy in diagnosing and managing complicated tissue patterns. A public dataset of 5000 histological pictures from eight CRC tissue types were trained on Vision Transformer and Compact Convolutional Transformer to achieve 93.3% and 95% accuracy, respectively. A new method integrating YOLOv4+ASFF and Swin Transformer was introduced in (Pan et al.; 2024) to detect and classify renal incidentalomas in CT images, potentially improving early identification and clinical practice. The model was trained using 1485 images, including 705 benign and 780 malignant, formed through image augmentation from 990 training and 495 validation images. The Swin Transformer backbone-based YOLOv4+ASFF network enhanced renal incidentaloma detection accuracy while maintaining competitive inference speed and strong generalization across datasets and settings.

Zhao (2022) used DL to classify skin cancer lesions, comparing CNNs with ViTs. This study's novel approach to overcome dataset imbalance by performing median frequency balancing, data augmentation, and sample size enhancement was combined with a detailed comparison of CNNs and ViTs on the skin lesion dataset, HAM10000. The CNN models (VGGNet and ResNet) and ViT models (regular ViT and DeepViT with reattention module) were trained using 10,015 dermoscopic pictures from seven skin lesions in the HAM10000 dataset. While CNNs exceed ViTs in accuracy (93.26% vs 84.20%), both can accurately classify skin cancer lesions. It was suggested that CNNs and ViTs could be combined with transfer learning to handle datasets with small sample numbers. Hence, (Li et al.; 2022) proposed the Trans-ResNet, a new architecture that combines CNN and Transformer strengths to classify Alzheimer's disease using MRI data. Trans-ResNet solves the problem of limited sample sizes in neuroimaging datasets by pre-training on a large-scale dataset for brain age estimate to capture both local and global dependencies in brain MRI images. The design included a ResNet-18 CNN encoder for local feature extraction and a Transformer encoder for global context modeling. It was pretrained for brain age estimation on the UK Biobank dataset and fine-tuned for Alzheimer's disease classification on the ADNI and AIBL datasets Trans-ResNet's classification accuracy in Alzheimer's disease prediction was 93.85% on the ADNI dataset and 93.94% on the AIBL dataset, outperforming CNN-based approaches and CNN-ViT ensemble models.

2.2 Pneumonia Detection using CNN and ViT

The authors in (Sharma et al.; 2020) proposed various CNN frameworks for feature extraction to detect pneumonia due to its prevalence and mortality rate from the chest X-ray dataset, using both the original and augmented datasets to explore the impact of the dataset size on the model's performance. This study reiterated on the importance of training the models from scratch, without using pretrained models which can lead to overfitting and generalization. The results achieved show that the CNN model trained on the augmented dataset performed better in terms of both validation and test accuracy, asserting the fact that the dataset size can also be a factor in disease classification. A research was proposed by (Labhane et al.; 2020) for the early detection of pediatric pneumonia from chest X-ray images using CNNs and transfer learning process. This study utilized four different CNN frameworks: standard CNN, VGG16, VGG19 and Inception to train on the pediatric pneumonia dataset that comprised of 2972 normal and 2992 pnuemonia X-rays providing a balanced dataset for proper classification performance. The efficiency of using the pretrained CNN transfer learning models trained on the ImageNet dataset was also evaluated with an accuracy of 97% highlighting the case for transfer learning for pneumonia disease detection. Six different CNN models like ResNet50, LeNet, AlexNet, GoogLeNet, VGG16 and StrideNet were used in the research presented in (Militante et al.; 2020), thus offering a encompassing assessment of CNN architectures for pneumonia classification. The models were trained on the chest x-ray dataset comprising of 28000 images of 224x224 resolution with a learning rate of 1e-4 and using Adam Optimizer. The final findings indicate that LeNet and GoogLeNet achieved the highest accuracy of 98%, followed by VGG16 with 97% and ResNet50 performing badly with only 80% accuracy.

Singh et al. (2024) The study used a public dataset of chest X-ray images from Kaggle, with three classes: normal, pneumonia, and COVID-19. The ViT model achieved an accuracy of 97.61%, sensitivity of 95%, and specificity of 98% in detecting pneumonia from

chest X-ray images. The research presents a simple yet powerful model for pneumonia detection using ViTs on a relatively small dataset.The research presented by (Jalalifar and Sadeghi-Naini; 2022) focused on classifying normal and abnormalities like pneumonia or COVID-19 from the chest X-ray images on a relatively smaall dataset comprising of 763 images, with 197 for COVID-19, 117 for pneumonia, and 449 for normal cases, addressing the data scarcity problem in medical imaging analysis. The detection framework was built on a Data-efficient image Transformer (DeiT) that employed a teacher-student scheme for training, achieved a test accuracy of 92.2% for classifying chest X-ray images into three distinct classes: normal, pneumonia and COVID-19 with the performance in par with CNNs, asserting the potential of Transformers for Pneumonia Classification. This study focuses on COVID-19 screening using chest radiography (X-ray and CT images). The authors in (Mondal et al.; 2021) proposed xViTCOS for COVID-19 screening using chest X-rays and CT images. This study highights the xViTCOS model's performance over other COVID-19 detection techniques focusing only on the regions of interest (ROI) in the chest X-rays, leading to the accurate diagnosis and precise localization of the disease lung region. The research concluded with the proposition that xViTCOS can complement or used in tandem with RT-PCR tests for rapid prognosis of COVID-19 for automated analysis of disease severity. (Wang et al.; 2023) introduces PneuNet, that combines ResNet18 and multi-head attention network from ViT for COVID-19 pneumonia diagnosis. The study used a custom dataset collected from seven public repositories, totalling 33,920 chest X-ray images divided and label under the three categories of cases of COVID-19, normal pneumonia, and healthy. PneuNet achieved a 94.96% accuracy on the test set for the multiclass classification problem, and for binary classification, 99.30% accuracy outperforming other DL models.

2.3 MLOps in Healthcare

Iqbal et al. (2023) focuses on the importance of MLOps in the healthcare sector for the deployment and management of DL models for medical image analysis using conventional machine learning models and CNN. It emphasizes the importance of collaboration among different companies and resources, including as technologies, algorithms, scripts, libraries, and tools, to automate data pretreatment, digestion, training, validation, and production processes. This research (Kundu and Bilgaiyan; 2022a) presented a comprehensive insight into the Machine Learning Operations (MLOps) practices for biomedical image classification, considering the specific requirements and challenges related to the handling and processing of biomedical data. This covers the processes behind the successful classification of biomedical images such as image filtering and segmentation using DL models like CNN, U-Net. The MLOps tools used at each stage of the software development workflow for biomedical images that includes data acquisition, preprocessing, model training, and deployment. The research concluded that the automation of medical image segmentation and classification models can be improved further with the use of MLOps pipelines and adhering to the MLOps practices for scalable and deployable models.

Khattak et al. (2023) An extensive review of MLHOps for reliable, efficient, and ethical deployment and maintenance of ML models in healthcare was presented in this research work. Proper guidelines and ethical practices for developers and clinical persons to deploy and maintain their MLOps models that addresses the long-term monitoring, updating, and other ethical issues. It also presented the data sources, pipeline engineering, deployment, monitoring, updating models, and ethical considerations specific to healthcare

use cases. The paper concludes that to fully realize the potential of machine learning in healthcare, practical considerations must be standardized and specified in engineering pipelines, termed MLHOps. A more practical approach of MLOps implementation in the diagnosis of COVID-19 from chest X-ray images was presented in (Kundu and Bilgaiyan; 2022b) using an automated hyper-parameter tuning pipeline to enhance the accuracy of DL models. This approach overcomes the time-consuming process of hyper-parameter tuning, leading to improved model accuracy with minimal human indulgence. A pre-trained model, CheXNet was used in this study and open-source tools like Docker and Jenkins were employed to create a Continuous Integration (CI) pipeline. The proposed MLOps approach successfully automated the hyper-parameter tuning process, resulting in a deep learning model that achieved an accuracy of 97.03%. The system continuously retrains itself until the desired accuracy is reached, eliminating the need for trained personnel during the model re-training stage.

3 Methodology

The research methodology presents the MLHOps based Pneumonia Detection from Chest X-Rays using the proposed CNN and ViT-based deep learning models. The primary focus of the study is to determine the performance of the CNN and ViT models in identifying pneumonia and deploy the best performing model using MLOps practices. This section discusses in brief: the dataset, data preprocessing techniques, model selection and training, model evaluation and MLOps integration. Figure.1 presents block diagram for better understanding of the process behind the design methodology.

3.1 Dataset

The NIH Chest X-ray dataset [1] comprising 5,863 frontal-view chest radiographs with labeled predictions of pneumonia was used in this study. The images were selected from pediatric patients of one to five years old from the NIH Clinical Center over the period of 1992-2015. The dataset contains 1,583 (27.0%) images with pneumonia findings and 4,280 (73.0%) normal images making it an unbalanced dataset which may facilitate the need for data augmentation for more reliable predictions.

3.2 Data Preprocessing

From the patient information provided in the dataset, it is clear that the age distribution is skewed towards the younger patients and the gender ratio shows a predominant male patient population. Along with the imbalanced class distribution with 2.7 times more normal cases than pneumonia cases, which can result in bias towards the majority class, the role of preprocessing the data is significant in working with such a dataset.

All images from the dataset have 1024 x 1024 resolution, but the pixel value distribution shows varying intensities across the dataset, likely due to differences in imaging equipment and other image capturing techniques. This necessitates the need for normalization during preprocessing. The pixel values of the images are normalized in the range [0,1] to ensure consistent intensity distributions across the dataset. The images also need to be resampled to a lesser resolution to reduce the workload on the model

[1] https://www.kaggle.com/datasets/tolgadincer/labeled-chest-xray-images

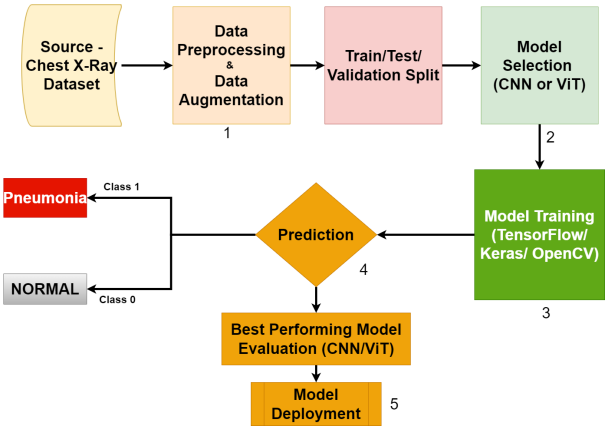


Figure 1: Pneumonia Detection with MLHOps.

improving the training and computing time without any adverse effects. Filtering methods can be used to remove any artifacts or labels from the images. Most importantly, data augmentation techniques like geometric transformations, intensity augmentations, random masking to better learn the representations etc., can be utilized to increase the diversity of the training samples and avoid overfitting of the model. These preprocessing and augmentation techniques are tailored carefully to improve the model's generalization ability and performance while avoiding insignificant transformations on the data. The augmented images can be combined with the original dataset to create the final training data for the ViT/CNN models.

3.3 Data Preparation

This is an important step in the model realization process where the preprocessed and augmented data is split into train, test and validation sets to evaluate the performance of the model. The training set comprises the majority of the data and is used to train the CNN and ViT models by optimizing its weights and biases. The validation set is the portion of data that is kept aside from training data for tuning the model hyperparameters and training process evaluation. The test set is also the section of data set aside to provide an unbiased estimation of the model's performance on the new/untrained data. The use of separate test set ensures that the model is not overtrained on any specific data, thereby avoiding any biases. The performance metrics like accuracy, precision, sensitivity and specificity can be used for evaluating the test set. As a general case, the train/validation/test data splits are in the ratio 75%, 15% and 15%, respectively.

3.4 CNN Model Architecture

CNN based frameworks have been used for medical image classification problems in the past decade, including pneumonia detection. In this study, we propose to use a custom CNN architecture with three blocks, with each block comprising of 2D convolutional layer with 16 filters, and a 3x3 kernel, followed by batch normalization, ReLU activation, MaxPooling and dropout. We also add a flattening layer to convert the feature maps into 1D vector. The output layer is a fully-connected layer (dense) with a single unit and sigmoid activation which is representative of a binary classification task. The CNN is trained using the binary cross-entropy loss function and the Adam optimizer. The number

of layers, filters, and its hyperparameters can be modified or fine-tuned to optimize the model performance.

3.5 ViT Model Architecture

Recently, Vision Transformers (ViT) have gained popularity over CNN for their ability to capture the global context and far-reaching dependencies in images, making it a natural choice for working Chest X-Ray images in this work. This could be beneficial for detecting subtle abnormalities indicative of pneumonia. The proposed ViT model uses the standard approach by splitting the input image into non-overlapping 16x16 size patches and linearly embeds them into a sequence of tokens of 768 dimensional vectors. This process is called patch embedding and a learnable class token is added to the sequence that serves as the representation for the image during classification. These sequence of patch embeddings and class tokens are then processed by a standard transformer encoder, having a multi-headed self-attention module and feed forward network to capture the relationships between different patches. The output of the transformer encoder is passed through a multi-layer perceptron (MLP) classification head to predict the probability of the presence or absence of pneumonia. The categorical cross-entropy function was the selected loss function and the optimizer was Adam. The use of Azure ML pipelines to create workflows that are modular, scalable and fast deployable in nature for training the registered, well-performing model on newer datasets can facilitate the process of ML development and deployment at faster speeds, leading to more reusable models in diverse application domains. The pipeline stages can be refined further to include more stages with complex hyperparameter optimizations and model evaluation processes. However, the focus of this research is limited to the deployment of the best-performing classification model using Azure ML pipelines in addition to other Azure services like Azure Compute, AzureML core, Azure Blob storage and others.

3.6 MLHops Pipeline

The entire MLHOps pipeline is implemented using Azure Machine Learning (Azure ML) services, that offers services for data versioning, model training on the cloud, model management, deployment, and monitoring. The following is the breakdown of the key components and processes in the MLHOps pipeline, as in Figure 1.

1. Data preprocessing and Data augmentation

2. Model Training (Python Data Science libraries like TensorFlow, Keras and Computer Vision library, OpenCV)

3. Model Prediction, where the predictions on the test data classifies the X-ray images as Normal (Class 0) and Pneumonia (Class 1). The Best-Performing Model Selection, where the performance of both the CNN and ViT models is evaluated using evaluation metrics like accuracy, sensitivity, specificity etc., and the overall best-performing model is selected for deployment.

4. The chosen best-performing model is then deployed to the production environment where it can serve for further predictions on other similar chest X-ray datasets from other sources.

4 Design Specification

The proposed pneumonia detection method classifies chest X-ray images as normal or pneumonia using CNN and ViT deep learning architectures. The model architecture of CNN and ViT model implementations are presented in this section.

4.1 Proposed CNN model

The chest radiographs are preprocessed by downsampling them to a standard 224 by 224 pixels and rescaling the intensity values to fall between 0 and 1. The proposed CNN architecture (see Figure 2) is structured into three primary convolutional stages succeeded by a densely connected classification module. Each convolutional stage is composed of a 2D convolution operation with a 3 by 3 spatial filter, followed by batch normalization, rectified linear unit (ReLU) nonlinearity, max pooling for downsampling, and dropout regularization. The series of convolutional stages progressively learn a hierarchy of discriminative visual patterns from the input radiographs. The densely connected head takes the final feature maps, reshapes them into a 1D vector representation, and feeds them through fully-connected layers with ReLU activations to perform the classification. The final output node applies a sigmoid activation to estimate the probability of pneumonia presence versus a normal case. The entire CNN is optimized end-to-end on the NIH Chest X-ray collection by minimizing the binary cross-entropy objective using the Adam stochastic gradient descent algorithm. To synthetically expand the training data and boost the model's invariance properties, random flips, rotations, zooms and translations are applied as data augmentations during the learning process.

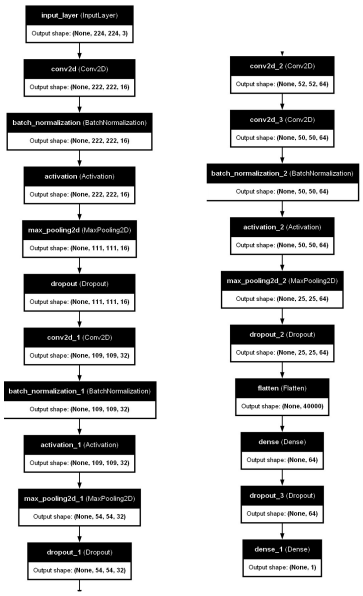
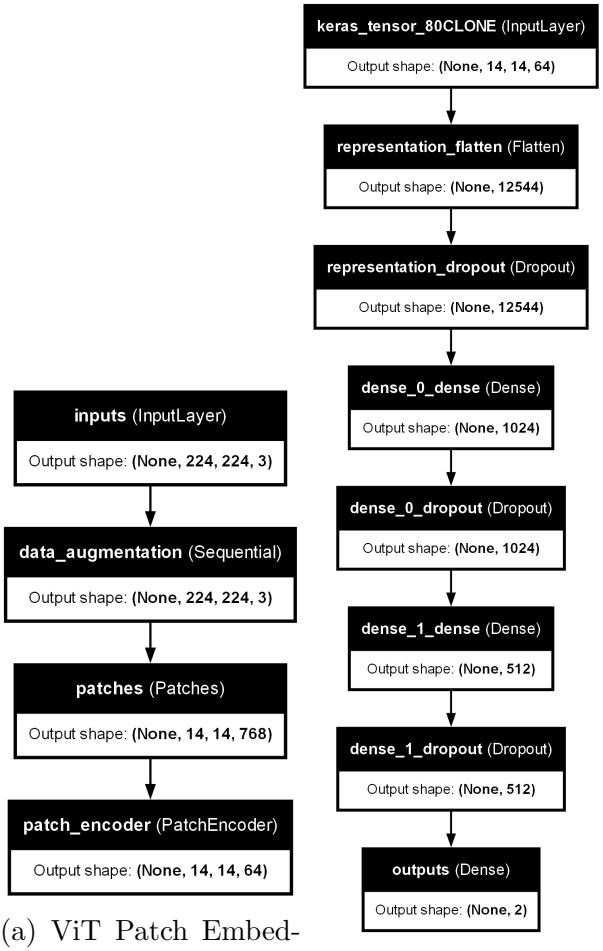


Figure 2: Proposed CNN architecture

4.2 Proposed ViT model

The input chest X-ray images are split into non-overlapping 16x16 patches and linearly embedded into a sequence of tokens of 768-dimensional vectors. This process, called patch embedding (See Figure 3 (a)), also adds a learnable class token to the sequence to serve

as the aggregate representation for classification. The sequence of patch embeddings and class token are processed by a standard transformer encoder, consisting of alternating layers of multi-head self-attention and feed-forward networks. The self-attention mechanism enables modeling of global dependencies between different regions of the image. Stochastic transformations like flipping, rotation, and zooming are performed by this layer on input images. These improvements help the model generalize by representing a variety of situations. The output shape is (None, 224, 224, 3), unchanged from the input. Here, 'None' means the batch size might vary. The Patches layer splits enhanced images into non-overlapping patches. The shape is (None, 14, 14, 768), suggesting that each image is divided into 14x14 pieces and compressed into a 768-dimensional vector space. Positional information from the patch encoder layer helps the model understand the sequence and position of each patch in the original image. The form (None, 14, 14, 64) shows that each patch is now a 64-dimensional vector. A thick PatchEncoder layer may reduce dimensionality from 768 to 64.



(a) ViT Patch Embedding

(b) ViT Classification

Figure 3: ViT Architecture for Pneumonia Detection

The output of the transformer encoder (Figure 3 (b)) corresponding to the class token is passed through a multi-layer perceptron (MLP) head to predict the probability of pneumonia vs normal. The classification head is typically the final part of a neural network model, responsible for taking the high-level features extracted by the previous layers and producing the final output or classification. In this case, the classification head

takes the flattened representation from the previous layers, applies two dense layers with dropout regularization, and produces a final output with two units, likely representing the probability or logits for the two classes, pneunomia, and normal.The ViT model is trained end-to-end on the NIH Chest X-ray dataset using the binary cross-entropy loss function and Adam optimizer. Data augmentation techniques like random cropping, horizontal flipping, and intensity are applied during training to improve robustness.

5 Implementation

5.1 Pneumonia Detection using CNN

The pneumonia detection system was implemented in an iterative manner following the design specifications. The final model and pipeline consisted of: Data Preprocessing: The NIH Chest X-ray dataset was version controlled and split into train/validation/test sets. Images were resized to 224x224 resolution and pixel values normalized. Data augmentation was implemented using Keras' ImageDataGenerator. Model Architecture: A custom CNN model with three convolutional blocks and a dense head was implemented using Keras Sequential API as shown in Figure 2. The model was compiled with binary cross-entropy loss and Adam optimizer. CNN Model Architecture is discussed in Section 3.4. The model was trained for 50 epochs on Azure ML compute clusters with early stopping and learning rate reduction callbacks. The binary cross-entropy loss and accuracy metrics were monitored on the validation set. The trained model was evaluated on the test set based on metrics like accuracy, and loss. Learning curves for the training and validation sets were plotted to analyze performance. The MLHOps process was divided into a three stage pipeline: data preprocessing, CNN Model training and Model Evaluation. The best-performing model based on validation accuracy was registered in the model registry. The training parameters for the CNN model is provided in Table 1.

Table 1: CNN Training Parameters

Parameter	CNN
Input Size	224x224
Convolutional Filters	[16,32,64]
Kernel Size	3x3
Dense Units	64
Learning Rate	0.000003
Dropout	[0.2, 0.2, 0.4, 0.5]

5.2 ViT based Pneumonia Detection

The preprocessing of NIH Chest X-ray dataset was done by resizing the images to 224x224 resolution, and normalizing the pixel values. A ViT-Base model with 12 transformer layers, 768 hidden dimensions, and 12 attention heads was implemented using TensorFlow and Keras. The MLP head consisted of two fully-connected layers. Pretrained weights from ImageNet were used to initialize the model. The model classification process is presented in Figure 3(b). The model was trained for 100 epochs with early stopping

to stop the training process if the validation loss does not decrease for more than 10 epochs on Azure ML compute clusters. The categorical cross-entropy was chosen as the loss function and Adam was the optimizer with a learning rate of 0.0001. The trained model was evaluated on the test set based on metrics like accuracy, precision, recall and f1-score. Visualizations like confusion matrices, accuracy and loss plots were logged. The best-performing model based on validation AUC was registered in the model registry. The model training parameters are presented in Table 2. The MLHOps process was divided into a three-stage pipeline: data preprocessing, ViT Model training and Model Evaluation. The best-performing model based on validation accuracy was registered in the model registry.

Table 2: ViT Training Parameters

Parameter	ViT
Patch Size	16x16
Project Dimension	64
Transformer Layers	8
Attention Heads	4
MLP units	[1024, 512]
Learning Rate	0.001
Weight Decay	0.0001

5.3 MLHops Implementation on Azure

The AzureML pipeline implementation for pneumonia detection is presented as a three-stage pipeline involving data preprocessing, ViT/CNN model training, and model evaluation. The first step of the process involves downloading the dataset from the Azure Blob storage using the right credentials that includes blob datastore name, account name, and account key. A compute cluster is created and assigned to each pipeline step. Since it is a sequential pipeline execution, the same compute instance can be utilized by the other pipelines too once a pipeline is completed. An execution environment with the necessary libraries is created for the pipeline as it runs in a virtualized environment ,and the libraries that are available in the main pipeline code are not available to the individual pipeline code files. The pipeline step is defined by the PythonScriptStep() function that takes the python script to run the pipeline as input with the input and output arguments. This enables moving the data easily to the next pipeline stage once the data has been processed and the pipeline step is executed. The CPU compute cluster and the coding environment is also defined to help with the execution of the pipeline. The order of execution of the pipeline is provided using the Pipeline() function. For instance, $pipeline = Pipeline(workspace, steps = [dataprepstep, modeltrainingstep, modelevaluationstep])$ will execute the dataprepstep pipeline first, followed by modeltrainingstep and finally, modelevaluationstep. Lastly, the pipeline is submitted for execution using the function, $run = Experiment.submit(pipeline)$ and can be tracked using the link generated upon successful submission of the pipeline. The created pipeline and its execution is presented in Figure 4.

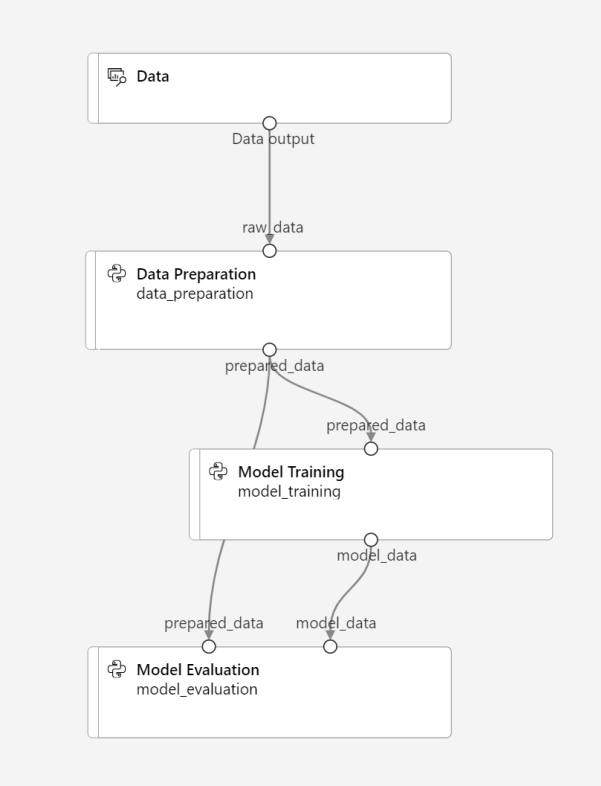


Figure 4: Three-Stage Pipeline Submitted for Execution

6 Evaluation

The CNN and ViT models were tested on NIH Chest X-ray images. The models' pneumonia vs. normal classification performance was measured using several parameters.

6.1 Experiment: CNN Model Evaluation Plots

The CNN model was trained with a learning rate of 3e-5, batch size of 32, for 50 epochs with an early stopping criteria if the validation loss does not reduce over 10 iterations of model training. Adam was chosen as the optimizer and binary cross entropy as the loss function. The model was able to achieve a validation accuracy of 0.9484 and test accuracy of 0.8605 which is good for the pneumonia detection classification problem. This section presents the evaluation results of the CNN model in terms of model training accuracy and loss plots (see Figure 5, 6). A classification report is also presented in Table-3 with metrics like precision, recall, and F1-score for better model performance analysis.

Table 3: CNN Evaluation Metrics

Classification	Precision	Recall	F1-score	Support
NORMAL	0.94	0.67	0.78	234
PNEUMONIA	0.83	0.97	0.90	390

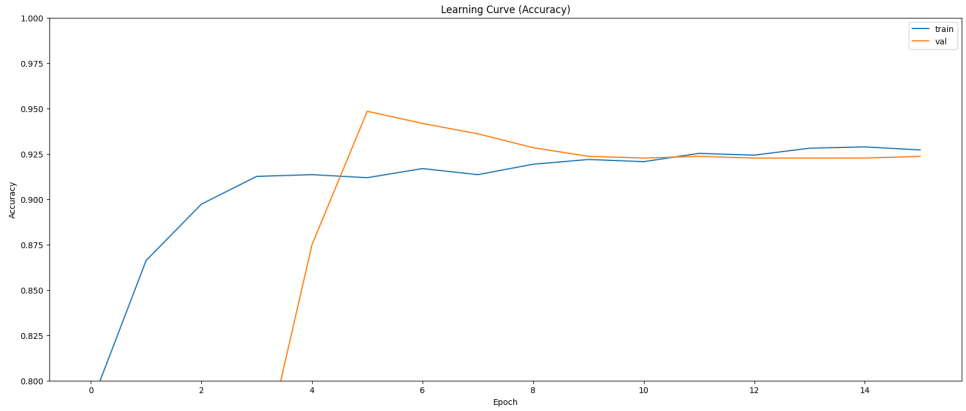


Figure 5: CNN Model Accuracy - Training & validation

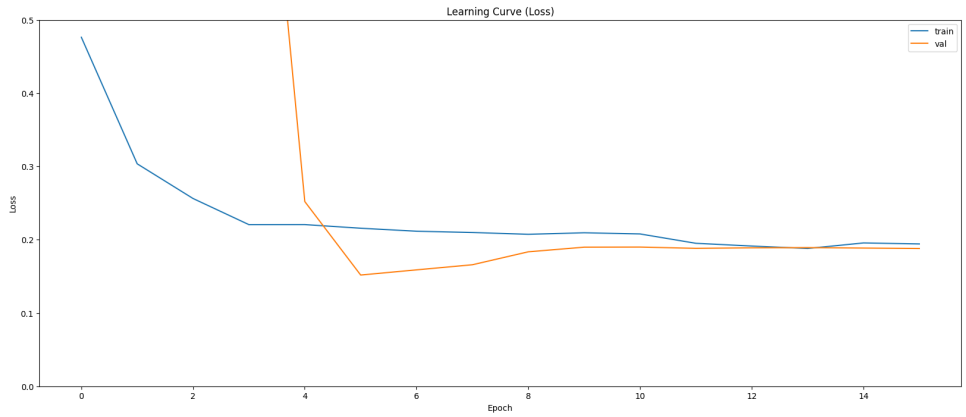


Figure 6: CNN Model loss - Training & validation

6.2 Experiment: ViT Model Evaluation Plots

The ViT model was trained with an epsilon value of 1e-6, learning rate of 0.001, weight decay of 0.0001, batch size of 32, for 40 epochs with an early stopping criteria if the validation area under curve (AUC) value does not reduce over 10 iterations of model training. Adam was chosen as the optimizer and categorical cross entropy as the loss function. The model was able to achieve a training accuracy of 0.9345 and validation accuracy of 0.9522 which is better than the CNN approach for the pneumonia detection classification problem. This section presents the evaluation results of the ViT model in terms of model training accuracy and loss plots (see Figure 7,8). A classification report is also presented in Table-4 with metrics like precision, recall, and F1-score for better model performance analysis.

Table 4: ViT Evaluation Metrics

Classification	Precision	Recall	F1-score	Support
NORMAL	0.91	0.58	0.71	234
PNEUMONIA	0.79	0.96	0.87	390

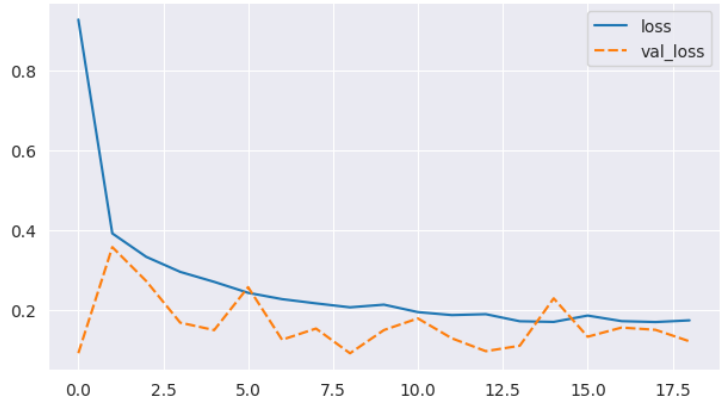


Figure 7: ViT Model loss - Training & validation

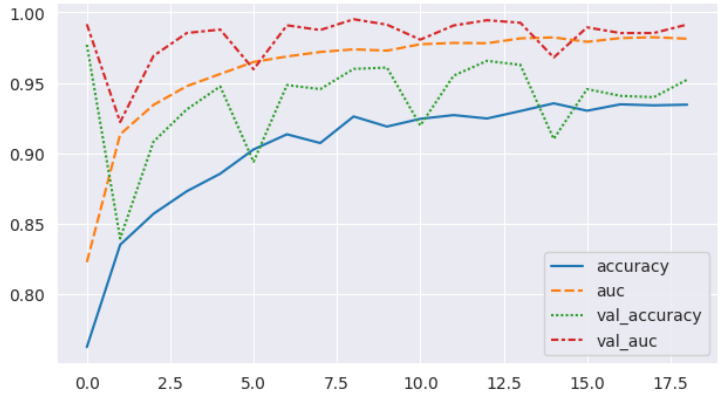


Figure 8: ViT AUC & Accuracy - Training & validation

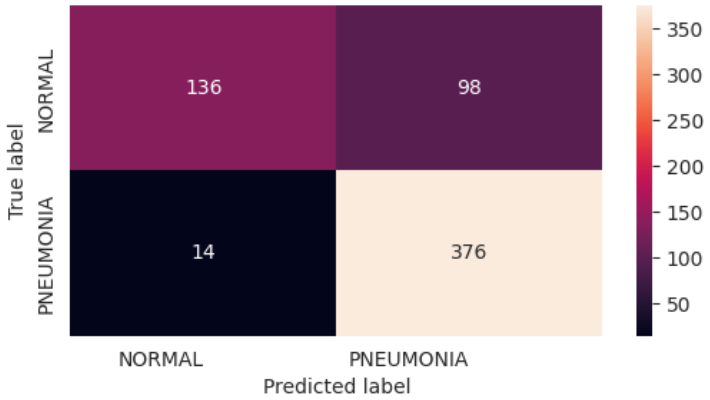


Figure 9: Confusion Matrix

6.3 Discussion

From a comparison of the results, it can be inferred that both the models perform well for the prediction of two classes, NORMAL and PNEUMONIA. This can be analyzed further from Table-3 and Table-4 where the classification report for both ViT and CNN model performance is presented. The CNN model performed improvedly well for the classification of the NORMAL class with a high recall value of 0.67 compared to 0.58 in ViT. This is important as recall scores can tell how many instances of a specific

class are correctly predicted out of all the actual instances of the class. CNN also had better balance in the predictions with an F1-score with a value of 0.78 over 0.71. But for the detection of PNEUMONIA class, both models performed exceptionally well with nothing much to differentiate between as seen in Table-3 and Table-4. However, the CNN model training and evaluation took about 7816 seconds while ViT completed the model evaluation in about 3018 seconds. This indicates an enormous difference in the speed of the model performance where ViT is more than twice faster than CNN in its detection performance. This is not particularly not preferrable for real-time deployment of these models using MLHOps where the CNN model's longer execution time can result in more cloud computing time and incur more costs in the long run. ViT with its shorter execution times and comparable performance metrics will be the ideal model to be deployed on the cloud.

7 Conclusion and Future Work

Vision Transformer (ViT) models constructed as end-to-end MLHOps pipelines on Azure were tested to see if they could outperform CNN models in pneumonia identification from chest X-ray images. The goals were to achieve improved classification results using an improved ViT-based DL approach, provide a comprehensive evaluation showing ViTs perform better than CNN models and implement the ML lifecycle using Azure ML pipelines. The research successfully addressed the question and achieved the objectives. Key findings of this research are: 1. The ViT model achieved a training accuracy of 93.45% and validation accuracy of 95.22%, over the CNN model (validation accuracy of 94.84% and test accuracy of 86.05%). The ViT model completed evaluation more than twice as fast as the CNN model (3018 seconds vs 7816 seconds), making it more suitable for real-time MLHOps deployment. MLHOps best practices were implemented using Azure ML services for data versioning, model training, and streamlined model deployment. Subsequent research should prioritize refining the findings to involve the classification of multiple diseases, optimizing the model deployment in various cloud environments, and integrating the models with medical information analytics.

References

Alsagheer, D., Xu, L. and Shi, W. (2023). Decentralized machine learning governance: Overview, opportunities, and challenges, *IEEE Access* .

An, F., Li, X. and Ma, X. (2021). Medical image classification algorithm based on visual attention mechanism-mcnn, *Oxidative Medicine and Cellular Longevity* **2021**: 1–12.

Ben Atitallah, S., Driss, M., Boulila, W., Koubaa, A. and Ben Ghezala, H. (2022). Fusion of convolutional neural networks based on dempster–shafer theory for automatic pneumonia detection from chest x-ray images, *International Journal of Imaging Systems and Technology* **32**(2): 658–672.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* .

Feng, N., Geng, X. and Qin, L. (2020). Study on mri medical image segmentation technology based on cnn-crf model, *IEEE Access* **8**: 60505–60514.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al. (2022). A survey on vision transformer, *IEEE transactions on pattern analysis and machine intelligence* **45**(1): 87–110.

Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E. K., Cohen-Adad, J. and Merhof, D. (2023). Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6202–6212.

Iori, M., Di Castelnuovo, C., Verzellesi, L., Meglioli, G., Lippolis, D. G., Nitrosi, A., Monelli, F., Besutti, G., Trojani, V., Bertolini, M. et al. (2022). Mortality prediction of covid-19 patients using radiomic and neural network features extracted from a wide chest x-ray sample size: A robust approach for different medical imbalanced scenarios, *Applied Sciences* **12**(8): 3903.

Iqbal, S., N. Qureshi, A., Li, J. and Mahmood, T. (2023). On the analyses of medical images using traditional machine learning techniques and convolutional neural networks, *Archives of Computational Methods in Engineering* **30**(5): 3173–3233.

Jafari, M., Auer, D., Francis, S., Garibaldi, J. and Chen, X. (2020). Dru-net: an efficient deep convolutional neural network for medical image segmentation, *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, IEEE, pp. 1144–1148.

Jalalifar, S. A. and Sadeghi-Naini, A. (2022). Data-efficient training of pure vision transformers for the task of chest x-ray abnormality detection using knowledge distillation, *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, pp. 1444–1447.

Khattak, F. K., Subasri, V., Krishnan, A., Dolatabadi, E., Pandya, D., Seyyed-Kalantari, L. and Rudzicz, F. (2023). Mlhops: Machine learning for healthcare operations, *arXiv preprint arXiv:2305.02474* .

Kundu, A. and Bilgaiyan, S. (2022a). *Automatic Enhancement of Deep Neural Networks for Diagnosis of COVID-19 Cases with X-ray Images Using MLOps*, pp. 155–165.

Kundu, A. and Bilgaiyan, S. (2022b). Automatic enhancement of deep neural networks for diagnosis of covid-19 cases with x-ray images using mlops, *Proceedings of Emerging Trends and Technologies on Intelligent Systems: ETTIS 2022*, Springer, pp. 155–165.

Labhane, G., Pansare, R., Maheshwari, S., Tiwari, R. and Shukla, A. (2020). Detection of pediatric pneumonia from chest x-ray images using cnn and transfer learning, *2020 3rd international conference on emerging technologies in computer engineering: machine learning and internet of things (ICETCE)*, IEEE, pp. 85–92.

Li, C., Cui, Y., Luo, N., Liu, Y., Bourgeat, P., Fripp, J. and Jiang, T. (2022). Transresnet: Integrating transformers and cnns for alzheimer's disease classification, *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, IEEE, pp. 1–5.

Li, X., Chen, F., Hao, H. and Li, M. (2020). A pneumonia detection method based on improved convolutional neural network, *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Vol. 1, IEEE, pp. 488–493.

Matsoukas, C., Haslum, J. F., Söderberg, M. and Smith, K. (2021). Is it time to replace cnns with transformers for medical images?, *arXiv preprint arXiv:2108.09038* .

Militante, S. V., Dionisio, N. V. and Sibbaluca, B. G. (2020). Pneumonia detection through adaptive deep learning models of convolutional neural networks, *2020 11th IEEE control and system graduate research colloquium (ICSGRC)*, IEEE, pp. 88–93.

Mondal, A. K., Bhattacharjee, A., Singla, P. and Prathosh, A. (2021). xvitcos: explainable vision transformer based covid-19 screening using radiography, *IEEE Journal of Translational Engineering in Health and Medicine* **10**: 1–10.

Pan, C., Chen, J. and Huang, R. (2024). Medical image detection and classification of renal incidentalomas based on yolov4+ asff swin transformer, *Journal of Radiation Research and Applied Sciences* **17**(2): 100845.

Parvaiz, A., Khalid, M. A., Zafar, R., Ameer, H., Ali, M. and Fraz, M. M. (2023). Vision transformers in medical computer vision—a contemplative retrospection, *Engineering Applications of Artificial Intelligence* **122**: 106126.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K. et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, *arXiv preprint arXiv:1711.05225* .

Sendak, M. P., Gao, M., Brajer, N. and Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels, *NPJ digital medicine* **3**(1): 41.

Sharma, H., Jain, J. S., Bansal, P. and Gupta, S. (2020). Feature extraction and classification of chest x-ray images using cnn to detect pneumonia, *2020 10th international conference on cloud computing, data science & engineering (Confluence)*, IEEE, pp. 227–231.

Singh, S., Kumar, M., Kumar, A., Verma, B. K., Abhishek, K. and Selvarajan, S. (2024). Efficient pneumonia detection using vision transformers on chest x-rays, *Scientific Reports* **14**(1): 2487.

Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V. and Hatamizadeh, A. (2022). Self-supervised pre-training of swin transformers for 3d medical image analysis, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740.

Wang, T., Nie, Z., Wang, R., Xu, Q., Huang, H., Xu, H., Xie, F. and Liu, X.-J. (2023). Pneunet: deep learning for covid-19 pneumonia diagnosis on chest x-ray image analysis using vision transformer, *Medical & Biological Engineering & Computing* pp. 1–14.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M. et al. (2019). Do no harm: a roadmap for responsible machine learning for health care, *Nature medicine* **25**(9): 1337–1340.

Zeid, M. A.-E., El-Bahnasy, K. and Abo-Youssef, S. (2021). Multiclass colorectal cancer histology images classification using vision transformers, *2021 tenth international conference on intelligent computing and information systems (ICICIS)*, IEEE, pp. 224–230.

Zhao, Z. (2022). Skin cancer classification based on convolutional neural networks and vision transformers, *Journal of Physics: Conference Series*, Vol. 2405, IOP Publishing, p. 012037.

Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L. and Yu, Y. (2023). nnformer: Volumetric medical image segmentation via a 3d transformer, *IEEE Transactions on Image Processing* .