Detection of Lung Cancer Using Vision Transformer

Rishi Sharma
Department of Computer Science
Rochester Institute of Technology
Rochester, United States
rs5501@rit.edu

***Abstract*—Lung cancer is a deadly disease, but if detected at the initial stages, the survival rate can be increased. Several diagnosis systems centered around neural network models have been used in the past to automate this process. Their results have been decent, but with the emergence of newer image classification techniques, there is a space for improvement. In this paper, we use Vision Transformer (ViT) as an image classification technique for the detection of lung cancer, with a Kaggle dataset consisting of a total of 15,000 lung histopathological images categorized into three different classes. With the power of the transformer model architecture and the presence of a multi-head self-attention layer, we expect a superior performance from ViT for this image classification task. Our final goal is to explore the effectiveness of ViT for the detection of lung cancer and determine if it outperforms pretrained Convolutional Neural Network (CNN) models in this area.**

***Index Terms*—Lung cancer detection, Vision Transformer**

I. Introduction

Lung cancer is a widespread and fatal disease caused by an abnormal growth of harmful cells in the lungs. Its case rates and number of deaths are increasing significantly every year, with lung cancer holding the highest percentage of cancer deaths in the United States [1]. However, the survival rate can be increased if it is diagnosed at an early stage. Generally, the diagnosis of lung cancer is done by the doctor using CT scans and X-rays, which is both time-consuming and error-prone. Therefore, an automated diagnosis system that detects lung cancer with a low misdiagnosis rate in practical settings would be beneficial.

In the past, Convolutional Neural Network (CNN) models have been used for automated detection of lung cancer. As a powerful image classification technique, CNNs can effectively capture localized and complex hierarchical features in image data. However, they lack the capability to capture the global relationships between those localized features, which could further improve accuracy. In recent times, transformer-based models (e.g., Vision Transformer) have emerged as an alternative to CNN models and have gained popularity. They have been shown to achieve comparable or even better performance than CNNs when pre-trained on large datasets (Dosovitskiy et al. [2]). Given the critical nature of lung cancer, even a slight improvement in model accuracy can potentially save lives. Therefore, there is a need for a more accurate model than CNNs to effectively detect the presence of cancer from lung tissue images.

The recent works we explored used methods such as deep CNN models, contextual feature extraction, auto-encoders, capsule networks, and reinforcement learning for lung cancer detection. Among these, the two notable ones that inspire our project are the diagnosis system for lung cancer detection based on CNN (Han et al. [3]), which showed the effectiveness of deep CNN models for lung cancer detection, and CA-Net (Liu et al. [4]), which highlighted the importance of contextual features by using a fusion of both lung nodule and contextual features as the basis for classification, which was mostly neglected in the traditional approaches. Based on our learning from these two works, we aim to use a model that surpasses CNNs in power and is capable of leveraging contextual (surrounding) features to achieve a higher accuracy in lung cancer detection.

In this paper, we choose Vision Transformer (ViT) as an effective image classification approach for lung cancer detection. ViT is a deep learning model that applies transformer architecture to computer vision tasks. Since it uses a multi-head self-attention layer (which is not used by CNN models), it can extract additional relationships between localized features (contextual features) that a CNN cannot capture, thereby making it the right choice of model. At a high level, ViT works by splitting the image into smaller patches, linearly projecting them, adding position and class embeddings, and then feeding them into transformer encoder layers.

Our proposed approach consists of multiple phases: (i) data preprocessing (e.g., image resizing, rescaling, and normalization), (ii) implementing a sequential model that uses a pre-trained ViT model as a layer with additional layers (e.g., flatten, batch normalization, dropout, and dense layers) on top of it, (iii) training the model along with hyperparameter tuning, (iv) testing the model with a test dataset, and (v) evaluating the performance metrics. We then compare the results with a baseline CNN model (e.g., ResNet50) using metrics and plots to determine if our approach outperforms CNN models for lung cancer detection.

II. Related Work

In this section, we explore some recent works focused on lung cancer detection.

F. Han et al. [1] proposed a Deep Convolutional Neural Network (DCNN) to identify and distinguish tumor and normal tissues in 5000 digitized lung tissue slide images that were captured at different magnifications (10X, 20X, 40X).

The DCNN architecture used in the study was the ResNet architecture as it was easier to optimize, and accuracy could be increased with the addition of more layers. Given the input images, parameters such as brightness, saturation, hue, etc. of the images were modified, and to enhance the randomization of the training set, a random dithering process was used. During the training process of the RestNet model, a random initialization method was used to set up the weights. The efficiency of the ResNet-18 model was then compared with RestNet-50, and it was found that they performed nearly similar, with ResNet-18 model achieving a slightly higher Area Under Curve (AUC) i.e. 0.999 compared to ResNet-50 model AUC (0.998), but the former model took half the time of the later model during the training and testing phase, hence using the ResNet-18 model proved sufficient for the dataset used. Given, the number of slicing samples to train on, the hardware used (computer with two 16-core 3.2GHz processors and two RTX2080Ti GPUs) and a large number of trainable parameters, it took a long time for training, but achieved a high accuracy. Additionally, the impact of digital image magnification (10X, 20X, 40X) on performance and efficiency was also assessed showing that the AUC performance didn't exceed significantly with high image magnification (20X, 40X), whereas the processing time and memory requirement were much less (1/16) for low magnification (10X), making it the suitable choice for this problem. The trained model was deployed in a digital slide scanner in a hospital and proved to show decent accuracy in lung cancer diagnosis.

Liu et al. [2] proposed a novel Context Attention Network (CA-Net) approach to lung cancer detection that emphasizes on the utilization of contextual features which were often ignored in the traditional approaches, originally centered around just nodule features such as shape and margin. CA-Net effectively extracted both nodule and contextual features from lung cancer images and fused them together before performing the lung nodule malignancy classification test. An attention-based mechanism was used to find the contextual features effectively and accurately by leveraging the nodule information as a reference. Given an input image, all the cropped nodule images were obtained using nodule detection, a 3D-Unet model was used as feature extractor to find the nodule features, and a Contextual Attention module was used for extracting surrounding contextual features for every individual nodule image. Following that, the feature maps were fused and run through a logistic classifier to get individual probabilities, that were finally combined to achieve the malignancy probability for the input image. The model was trained on 130 epochs, and an SGD optimizer was used with a learning rate of 0.01 to achieve the best results. The performance of the proposed model was evaluated on Data Science Bowl (DSB) 2017 lung cancer dataset provided in a competition on Kaggle containing labeled images for 2101 patients, and the results showed that it outperformed the winner by 2.5% on AUC, thereby concluding that accurately capturing the contextual features can give better accuracy in lung cancer detection.

Cai et al. [3] proposed a blood-based method named DeepMeth using circulating tumor DNA (ctDNA) data for lung cancer detection that is both non-invasive and efficient in terms of cost compared to manual clinical diagnosis or CT Scans. Being the first study to represent methylation regions by continuous vectors, it effectively captured early signs of the disease that were generally ignored by earlier methods as they didn't focus on methylation positions and patterns. The framework of DeepMeth was divided into 2 phases, involving learning about the methylation region representations/vectors using a residual-based auto-encoder model and passing these learned region vectors through a widely adopted classifier to obtain the final prediction determining if the input sample is malignant or benign. Given the high cost of collecting ctDNA methylation samples, there were fewer samples, but each sample had a high count of regions and reads, which could potentially lead to a dimensionality problem. As previous studies had shown a very weak correlation among these regions, hence for the sake of simplicity, the proposed method considered each region independently. The encoder process of the autoencoder started with 7X7 convolution filter followed by 4 residual blocks to get the region vector. Each residual block consisted of a pair of 3X3 convolution filters, Batch Normalization layers, and ReLU layers in a specific order. With the region vector as input, the decoder process followed reversed data flow, and the use of transposed convolution filters to reconstruct the region. DeepMeth was tested and evaluated with 6 different classifiers, of which three were ensemble classifiers (Random Forest, XGBoost, LightGBM), and the other three were deep learning models (MLP, CNN, and recurrent neural network (RNN)). Comparing its performance with the 4 state-of-the-art baseline methods (evaluated with the same classifiers), it was observed that DeepMeth outperformed them by around by 5%-8% in terms of AUC, and the results showed that combining DeepMeth with the baseline metrics can further enhance the prediction performance. The effect of the size of region vectors was also studied, and the result showed that performance first increased with an increase in size, and decreased after a certain size, with the maximum accuracy at size=10. As stated in the paper, DeepMeth has been deployed in more than 94 hospitals for clinical diagnosis and has been performing well.

The study by Mobiny et al. [4] explored the effectiveness of capsule networks (CapsNets) in lung cancer diagnosis. Presented that CapsNets could overcome the drawbacks of CNNs related to discarding important features like location, direction of objects in images, and its need for a large number of training samples to improve accuracy and avoid overfitting. CapsNets were known to be computationally expensive and time-consuming to train. To address this, a Fast Capsule Network (FastCapsNet) was proposed, which modified the dynamic routing mechanism used by the original CapsNets, making it computationally more efficient. Permitting only a single capsule at each pixel location in the image, meant all the capsules in the PrimaryCaps Layer having the same pixel location would have the same routing coefficient, thereby reducing the number of routing coefficients by a factor of 32, and highly improving the speed and efficiency of the

network. The 2D and 3D versions of the proposed FastCapsNet were trained and tested on 7400 images generated from 226 unique CT scans along with respective versions of original CapsNet and CNN models (AlexNet, ResNet). Performance comparison showed that the proposed FastNet model was 3 times faster than the original CapsNets with the same accuracy. It also significantly outperformed the CNN models in terms of accuracy when the number of samples is less and produced comparable results otherwise.

Wang et al. [5] presented a unique idea of formulating the stochastic nature and history of lung cancer as a Partially Observable Markov Decision Process (POMDP), changing it to a fully observable belief MDP through the application of a theorem, and finally proving the lung cancer diagnosis problem to be equivalent to the Markov optimal stopping problem. Furthermore, a reinforcement learning-based approach named EarlyStop-RL was proposed to find a solution (interpretable stopping rule) for the optimal stopping problem by utilizing the Snell envelope structure. The EarlyStop-RL algorithm consisted of a belief update phase that involved updating the belief probabilistic values for patients based on past observations after every medical check-up, and an optimal stopping phase to decide if a patient is negative, positive, or needs further checkups using current belief values. A cost model was proposed to take into consideration the individual risks of false positives/negatives & delayed diagnosis, aiming to minimize overall cost to find the perfect balance between misdiagnosis and delayed diagnosis. The algorithm was evaluated on a clinical dataset (the National Lung Screening Trial) consisting of scans from 1951 patients, and its performance was compared with two clinically adopted models (Lung-RADS, Brock model), and the Google AI model (recently popular) based on multiple parameters. Results showed that the EarlyStop-RL algorithm outperformed the other 3 methods based on the 6 performance metrics used such as false-positive/negative rate, early diagnosis rate, etc. The limitation of the algorithm was considered to be because of the lack of features, and the intent to improve by increasing the number of features along with the use of more powerful networks for the observation model was stated.

III. Methodology

In this section, we discuss in detail about the dataset used and the data preprocessing steps, followed by an in-depth description of the ViT architecture and fine-tuning process for the baseline pretrained ResNet50 and ViT-based models, and their training process.

A. Dataset and Preprocessing

For this study, we used a Kaggle Dataset [8] consisting of 15,000 labeled lung histopathological images, categorized into 3 classes:

(i) Lung benign tissue (lung_n)
(ii) Lung adenocarcinoma (lung_aca)
(iii) Lung squamous cell carcinoma (lung_scc)

The dataset consists of 5,000 images for each class, and each image is 786 x 786 pixels in JPEG format.

After loading the dataset into a pandas dataframe, multiple preprocessing techniques were applied to prepare the data to be fed to the models for lung cancer detection. Lung tissue images were first resized to a suitable size format (224 x 224 pixels) for input to CNN (ResNet50) and ViT-based models. The pixel values of the images were rescaled from [0, 255] to [0, 1], to better stabilize the training process of the models, eliminate the bias towards a particular range of pixel values, and to avoid the problem of exploding gradients. Following this, we performed a train-validation-test split of 80%-10%-10% on the original dataset for training and testing purposes and one-hot encoded all the class labels. The train, validation and test samples were then randomly shuffled to enhance data generalization for improved learning.

B. ViT Architecture

ViT was first proposed by Dosovitskiy et al. [2] and is based on transformer architecture that was originally developed to be used for NLP problems. ViT applies that architecture to computer vision tasks and has been known to achieve comparable or even better performance than state of the art CNN models when pretrained on large datasets.

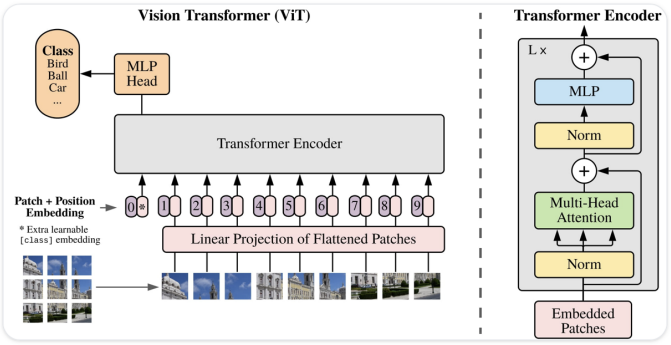

Fig. 1. High Level ViT Model architecture.

Figure 1. provides a high level diagram for the workflow and important components that a ViT model comprises of.

First, the input image is split into fixed size blocks called patches, and are linearly embedded with position embeddings, that store position information for every patch which will be beneficial to learn relationships between patches. An extra learnable class embedding parameter is prepended to the embedding vector, and is learned through back-propagation during training to be further used by the MLP head for classification.

The embedding vector is then passed to a series of a transformer encoder blocks, with each block consisting of two components: a multi-head self-attention layer and a feed forward network. The normalization layer and residual connection are used to stabilize the training process and reduce the training time. The MLP head takes the output of the last transformer encoder block and the class embedding vector as input and is responsible for giving the final class outputs.

ViT provides powerful image processing capabilities, and we believe it can be a superior alternative to CNN for lung cancer detection. With the use of multi-head self-attention layers, the model can attend to multiple parts in the image simultaneously and can learn long-range features and global relationships in the images that a CNN can't capture.

C. Fine-tuned ResNet50 Model Architecture

To gaudge the effectiveness of the ViT model, we plan to compare its performance with CNN, which has been considered state-of-the-art for image processing tasks for many years.

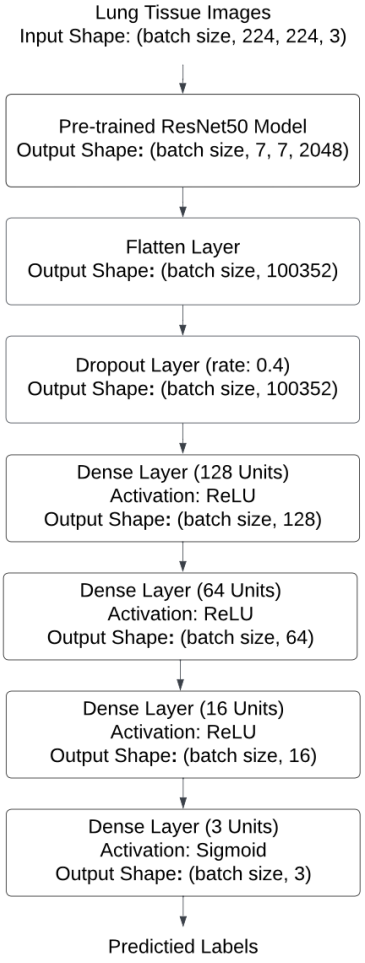

Fig. 2. Flowchart of Fine-Tuned ResNet50 model architecture.

From our literature review, we know that F. Han et al. [1] used a Deep CNN model based on ResNet50 architecture for lung cancer detection from lung tissue images. For our study, we use the ResNet50 model as a baseline for comparison with the performance of the ViT model.

ResNet50 is a CNN model that is 50 layers deep and is known to be computationally efficient when working with large datasets and deep neural networks. For our analysis, we used its pre-trained weights (trained on the ImageNet dataset) and fine-tuned the model to suit our requirements.

Fig. 2 illustrates a high level architecture of our fine-tuned ResNet50 model. It uses the pre-trained ResNet50 model as a base layer, followed by a flatten layer which is important to convert the [7, 7, 2048] sized output tensor from ResNet50 model into a 1D vector of size 1000352. A dropout layer is used to handle overfitting, followed by 4 fully-connected layers. The first 3 fully-connected layers aim to learn features that the pre-trained ResNet50 model couldn't capture. The final layer, with 3 units and a sigmoid activation function, is used for output classification.

Since, our dataset differs from the ImageNet dataset classes on which ResNet50 was pretrained, the addition of multiple dense layers was important to capture complex features, leading to a richer feature representation. Adjusting the dropout rate, and number of units in the dense layers also played a significant role to avoid model overfitting.

D. Fine-tuned ViT-based Model Architecture

Unlike CNN, ViT processes an image by dividing it into fixed-size patches instead of using convolutions. For our study, we used the Vit-Keras package [9] which provides a Keras implementation of the ViT model proposed in the original ViT paper (Dosovitskiy et al. [2]).

Fig. 3 shows a high level architecture of our fine-tuned ViT-based sequential model implementation.

It comprises of the ViT-Base (ViT_B32) model, pre-trained on the ImageNet-21k dataset, as the initial layer. The output of the pretrained ViT model is a 768 sized embedding vector. This is followed by a flatten layer to flatten the embedding vector, a dropout layer, and 4 fully-connected dense layers.

We used dropout layer with a rate of 0.2 to randomly drop 20% of the units from the previous layer, thereby helping in preventing overfitting. The three fully-connected layers (with units 64, 32 and 16 units, respectively, and ReLU activation function) aim to learn complex non-linear features from the output embedding vector of the last transformer encoder block. The final dense layer, with 3 units and a softmax activation function, acts as the output classification layer.

Given the more powerful transformer based architecture of the ViT-base model, it is capable to effectively capture complex features in the image through its self-attention mechanism. For this reason, we used less number of dense units while fine-tuning the ViT model compared to the ResNet50 model. The self attention mechanism in ViT provides built-in overfitting handling by focusing on the relevant features in the image and ignoring the less relevant ones, thereby improving model generalization. Thus, we used a lower dropout rate of 0.2 for fine-tuning the ViT compared to the ResNet50 model, which used a dropout rate of 0.4.

We used ViT-base model configuration with a patch size of 32, that contains a total of 12 transformer-encoder blocks. Each block containing a multi-head self-attention layer with 12 attention heads, normalization layers, and a Multi-Layer Perceptron (MLP) block.

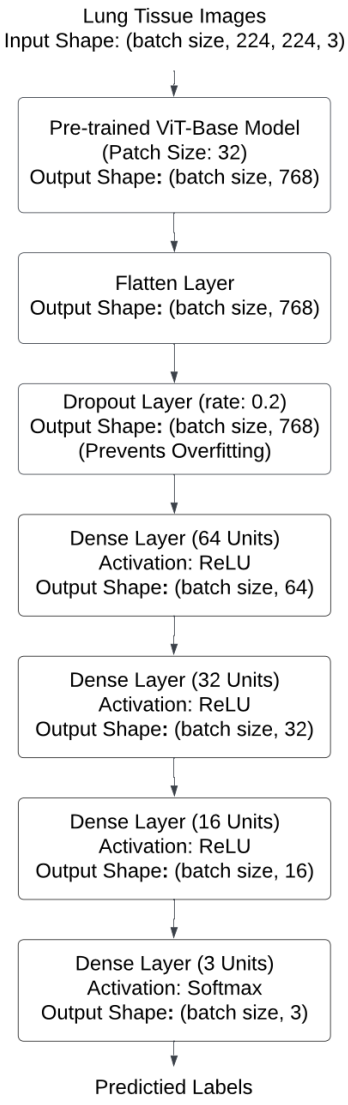

Fig. 3. Flowchart of Fine-Tuned ViT-based model architecture.

E. Training Process

During the training process, we compiled both the pretrained ResNet50 and the ViT-based model using the Adam optimizer, categorical crossentropy as the loss function, and a batch size of 64. Both the models were trained and validated for 8 epochs and converged properly. The ViT-based model, with over 87M trainable parameters, took longer to train (around 5-6 hours) compared to the ResNet50 model, which had 36M parameters and took 3 hours using the T4 GPU provided by Google Colab.

IV. Experimental Results

This section describes the training and testing results we achieved from the fine-runed ResNet50 and ViT-based model, using accuracy and loss as the evaluation metrics. Additionally, we discuss the methods followed to enhance the ViT model's accuracy and its performance comparison.

A. Performance of Pretrained ResNet50 Model

On the training data, it achieved 74.30% accuracy with a loss value of 0.6099, and on the validation data, it achieved 75.27% accuracy with a loss value of 0.5693.

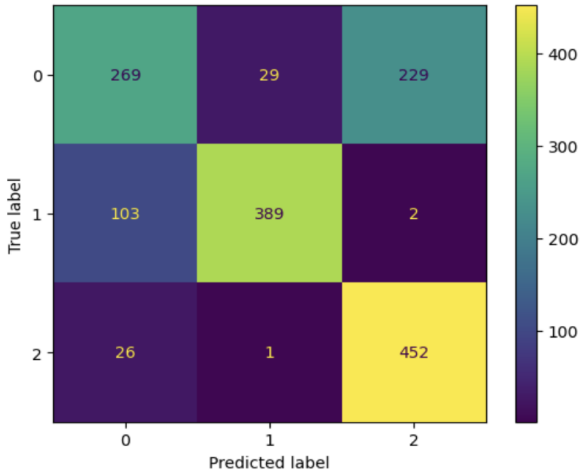

Fig. 4. Confusion Matrix of ResNet50 Model Test Performance.

However, when used for prediction on unseen test data, it achieved 74% accuracy, which is not very promising.

Fig 4. shows the number of correctly and misclassified predictions by the ResNet50 model. It can be observed that a lot of test samples are misclassified with maximum inaccuracies between class 0 (lung_aca) and class 2 (lung_scc) samples.

B. Experimentation Methods

We explored the following methods to improve the ViT model's accuracy:

(i) Train-validation-test splits: We experimented with different splits. The model with the 70%-15%-15% split didn't learn that well, giving a sub-optimal training and test accuracy of 85% and 82% respectively, possibly because of lesser data to train on. The 90%-5%-5% split led to a higher training accuracy of 95% because of more training data, but a reduced test accuracy of 78%, indicating overfitting.

We observed that the model generalized best with the original 80%-10%-10% split having sufficient training data (12,000 samples), as well as enough data to validate (1500 samples) and test the model (1500 samples).

(ii) Activation functions: Explored multiple activation functions, including ReLU, LeakyReLU, Sigmoid, and Softmax. We used ReLU activation function for the additional dense layers because it is computationally efficient, helps in learning non-linear features, and prevents the vanishing gradient problem, which can hinder the model's performance.

For the output classification layer, we used the Softmax activation function. It converts the output logits of the network into a probability distribution over classes, with each probability value representing the model's confidence in predicting each class. The sum of all probabilities equals 1, making it an excellent choice for multi-class classification problems.

(iii) Batch size tuning: We tried batch sizes of 32, 64, and 128. Batch size of 128 led to faster training with 94 steps in each epoch but reduced test accuracy (90%), possibly because of fewer gradient updates leading to a suboptimal solution. A batch size of 32 resulted in improved test accuracy (96%) due to more frequent gradient updates with 376 steps in each epoch, but was slower to train. Batch size 64 provided the best balance between training time and accuracy taking around half the time of batch size 32 to train, and achieving a similar test accuracy of 96%.

(iv) Model complexity: We experimented by adding and removing various Keras layers (self-attention, dense layers, dropout layers, and batch normalization layers) on top of the ViT-base model. Initially we added a self-attention layer (to further enhance contextual understanding) and a batch normalization layer layer (to stabilize training) along with a dropout layer and fully connected dense layers to fine-tune the ViT model. While it achieved a training accuracy of 97%, it didn't performed well on test data with an accuracy of 82% and thus resulted in overfitting.

We found that a comparatively simpler structure with a flatten layer, dropout layer, and four fully connected layers (with 64, 32, 16, and 3 units) on top of the pretrained ViT-Base model (as shown in Fig. 3) performed best on unseen test data.

(v) Overfitting Handling: Used dropout layer, adjusted dropout rate value, and modified the number of units in dense layers as regularization techniques to prevent overfitting during the model fine-tuning process.

(vi) Data Normalization: Applied mean and standard deviation normalization to the images during the data preprocessing step, setting the mean of all pixels to 0 and the standard deviation to 1. This helped the ViT model to converge faster, and generalize better.

C. *Performance of ViT-based Model*

Our fine-tuned ViT-based model performed much better, achieving a 97.01% accuracy on the training data and 94.87% on the validation data.

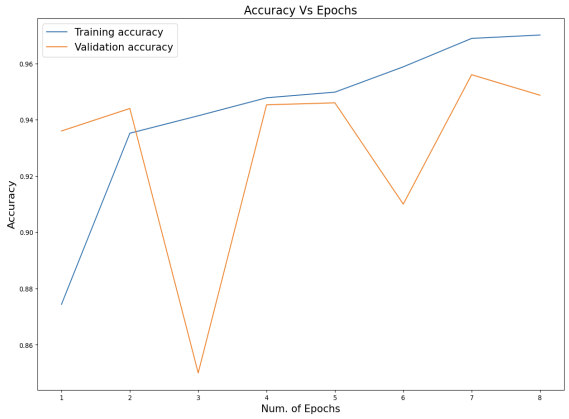

Fig. 5. Accuracy vs Epoch Plot.

The training and validation loss values were also improved, with values of 0.0755 and 0.1347, respectively.

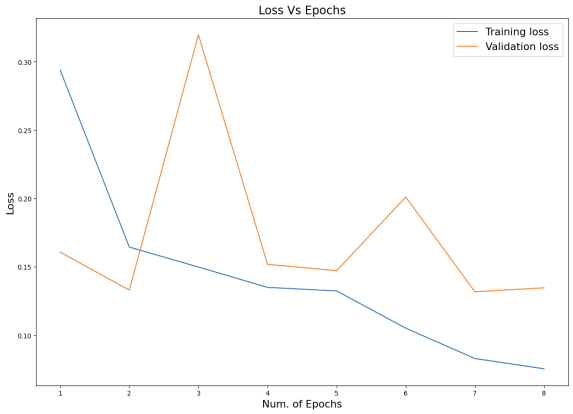

Fig. 6. Loss vs Epoch Plot.

Fig. 5 and 6 illustrate the training progress for our fine-tuned ViT model across each epoch.

We can observe a consistent upward trend in the training accuracy and a corresponding downward trend in training loss, representing that model learned training data progressively well with each epoch. The validation accuracy and loss show the same trend, but with fluctuation at epoch 3, possibly because of shuffling of data before splitting into batches in validation set. We shuffled the train and validation data before splitting them into batches for each epoch, in order to avoid any bias from the order of data points and improve generalization, and thus expected a few fluctuations, as seen here.

On the test data, it achieved a noteworthy 96% accuracy, leveraging the advantages of transformer architecture and self-attention mechanism to capture global relationships in images.

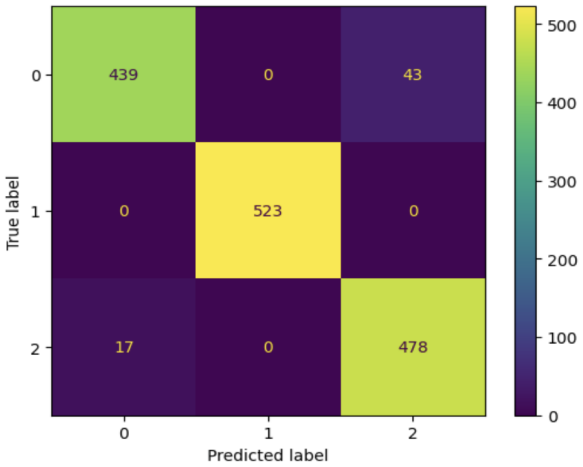

Fig. 7. Confusion Matrix of ViT-based Model Test Performance.

Fig 7. illustrates the prediction counts by our fine-tuned ViT model, indicating significantly less number of misclassified samples. The precision, recall and f1-score of 1.00 for class 1

(lung_n) suggests that the ViT model performed exceptionally well in predicting class 1, with zero false positives and false negatives.

The ResNet50 model faced a high number of misclassifications in distinguishing between Class 0 (lung_aca) and Class 1 (lung_scc) because both these classes are quite similar and have the presence of lung cancer, with only the cancer intensity varying. In this case, the ViT model performed significantly better with f-1 scores of 0.94 for both class 0 and class 2, thereby resulting in an overall 96% accuracy on the test data.

D. Model Performance Comparison

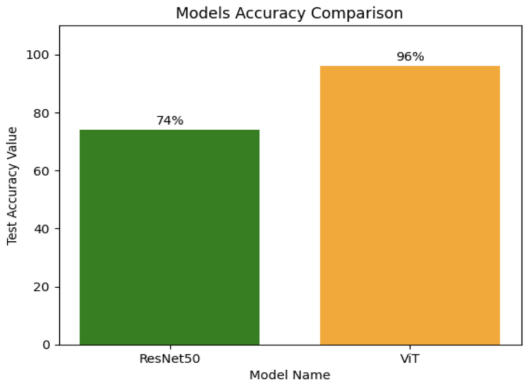

Fig. 8. Baseline ResNet50 vs ViT Model Test Performance.

Fig. 8 shows a bar chart comparing the test accuracy of ViT with the Baseline ResNet50 model.

We observed that the ViT model achieves a superior overall test accuracy (96%), with higher precision and recall values for all 3 classes, thereby outperforming the ResNet50 model in prediction accuracy.

The performance results based on accuracy and loss support our initial hypothesis that the ViT model based on a powerful transformer based architecture with its ability to use multi-head self attention layers, can extract long-range dependencies and contextual relationships in images that CNNs often miss, and these additional contextual features can further enhance accuracy, as observed in our case.

V. Conclusion/Future Work

We have conducted an analysis of our ViT-based sequential model, evaluated its performance, and observed that it outperforms the baseline pretrained ResNet50 model in accuracy for lung cancer detection with the dataset used.

For future work, we plan to experiment with different optimizers, such as RectifiedAdam, and apply label smoothing to the loss function. Label smoothing can help minimize overconfidence in predictions and improve generalization. Additionaly, we intend to apply our fine-tuned ViT model to other medical image diagnosis tasks, such as Pneumonia detection, Tumor diagnosis etc.

References

[1] American Cancer Society. (n.d.). https://www.cancer.org/cancer/types/lung-cancer/about/key-statistics.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,". arXiv preprint arXiv:2010.11929, 2020.

[3] F. Han, L. Yu and Y. Jiang, "Computer-aided diagnosis system of lung carcinoma using Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 2953-2958.

[4] M. Liu, F. Zhang, X. Sun, Y. Yu, and Y. Wang, "CA-Net: Leveraging Contextual Features for Lung Cancer Prediction,". In: de Bruijne, M., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science(), vol 12905.

[5] X. Cai, J. Tao, S. Wang, Z. Wang, J. Wang, M. Li, H. Wang, X. Tu, H. Yang, J.-B. Fan, and H. Ji, Noninvasive Lung Cancer Early Detection via Deep Methylation Representation Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 11828-11836, 2022.

[6] A. Mobiny and H. Van Nguyen, "Fast CapsNet for Lung Cancer Screening,". In: Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science(), vol 11071.

[7] Y. Wang, Q. Zhang, L. Ying, and C. Zhou, "Deep Reinforcement Learning for Early Diagnosis of Lung Cancer,". Proceedings of the AAAI Conference on Artificial Intelligence, 38(20), 22410-22419, 2024.

[8] Larxel (2020) Lung and colon cancer histopathological images, Kaggle. https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images.

[9] "vit-keras," PyPI, May 03, 2023. https://pypi.org/project/vit-keras/