Journal of High School Science

Deep learning for lung cancer detection and classification in CT scans

Pawar S

Submitted: November 21, 2023, Revised: version 1, January 16, 2024, version 2, January 18, 2024
Accepted: January 21, 2024

Abstract

Over the past few centuries, lung cancer has been the deadliest cancer for both men and women in the U.S., accounting for 12% of all new cancer diagnoses and 25% of all cancer deaths. To address this challenge, we developed a series of machine learning models that can be used to detect the presence and type of lung cancer based on an individual's CT scan. This article will discuss the selected model architectures and model performances. We found the highest-performing model to be a fine-tuned ResNet50, successfully predicting the presence and cancer type with a test accuracy of 88.89%. The implications of this project are to detect and diagnose lung cancer to aid doctors in making correct diagnoses of lung cancer, thus improving the chances of patient survival.

Keywords

Lung cancer, Deep learning, Transfer learning, Adenocarcinoma, Squamous cell carcinoma, Large cell carcinoma, CT scans, ResNet50, Data augmentation

Shreya Pawar, American High School, 33647 Pacheco Dr, Fremont, CA 94555, USA.
shreyapawar011@gmail.com

1 Introduction

1.1 Background

Lung cancer is the deadliest cancer worldwide, accounting for 12% of all new cancer diagnoses and 25% of all cancer deaths, making it the leading cause of cancer fatality (1). This project aims to aid doctors by classifying lung cancer variations using deep learning models. Deep learning is a subgroup of machine learning that focuses on making predictions by identifying patterns in data using neural networks. It has served as an assisting tool in cancer phenotyping and therapy for decades and has been widely implemented in advanced approaches for early detection, cancer type classification, signature extraction, tumor microenvironment, deconvolution, prognosis prediction, and drug response evaluation (2). In this project, we use deep learning to classify the CT scans into their respective classes. Specifically, we employ a series of convolutional neural networks (CNNs) to process and understand the image data. CNNs are networks for deep learning that learn patterns in images through the application of filters (3). The model uses these filters to analyze and categorize each class, and eventually output the predicted cancer variation. This project focuses on the diagnosis of three different types of lung cancer, which are compared to normal lung scans (Class 2). The first is Adenocarcinoma (Class 0), a type of cancer that forms in the glandular epithelial cells, which secrete mucus, digestive juices, or other fluids, starting in glands that line the insides of the organs, which is why it can affect different body areas. The survival rate for this variation is 32.3% (4). The second variation is Large Cell Carcinoma (Class 1), a type of cancer that can begin in several types of large cells, appearing larger than typical cancer cells when viewed under a microscope. The survival rate for this cancer type is 24.9% (5). The last type of lung cancer is Squamous Cell Carcinoma (Class 3), which develops in the squamous cells, which are thin, flat cells found in the skin and the tissue lining the lungs. The survival rate for this variation is 24% (6).

2 Materials and Methods

2.1 Dataset Description and Data Cleaning

2.1.1 *Dataset Attributes*

The dataset consists of 1,000 lung CT scans. Each image is a 256 x 256 RGB scan oriented along the axial plane, depicting the cross-section of each lung. The images are mapped by an x and y-axis. These scans are split into the following four classes: Adenocarcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma, and Normal. The dataset is further distributed into train, test, and validation sets, at a corresponding ratio of 7:2:1 with 613 training, 315 testing, and 72 validation images. For certain models, the images were augmented to increase accuracy. The augmentations used include rescaling, random flip, and random rotation. The dataset used for the project is available for download *via* Kaggle (7).

2.1.2 *Data Cleaning*

The raw data contained various pixel ranges and contrasts. Using the mean value, which is the sum of pixel values divided by the total number of pixel values, this wide range of variations was standardized through every scan. Specifically, the data was averaged such that it has a mean pixel value of 0 and a standard deviation of 1.

2.1.3 *Class Distribution*

The class distribution of slices for the entire dataset is as follows: Adenocarcinoma: 338 slices; Large Cell Carcinoma: 187 slices; Squamous Cell Carcinoma: 260 slices; Normal: 215 slices. During the train-test-validation split, the 1,000 scans were shuffled and distributed

into random-sized train, test, and validation sets. Due to this process, the class ratios for each set differ.

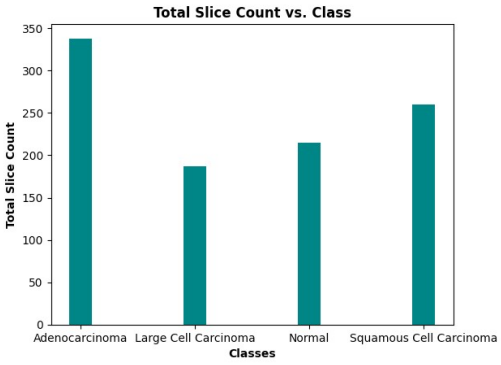

Figure 1. A bar plot depicting the slice count for each class

Figure 1 shows the slice count for each class and demonstrates a class imbalance, with nearly double the amount of Adenocarcinoma images as there is Large Cell Carcinoma. This discrepancy may cause the models to be less accurate when classifying Large Cell Carcinoma scans while being more accurate when categorizing the Adenocarcinoma cancer images.

2.2 Exploratory Dataset Analysis

2.2.1 *Class Visualization*

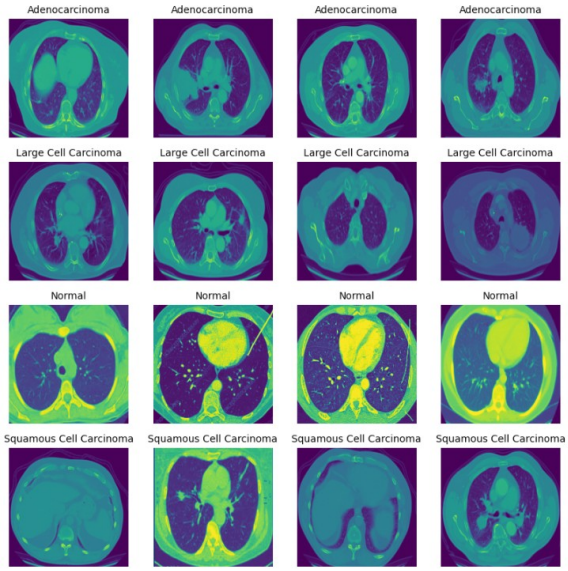

Figure 2. A visual representation of CT scans from each of the four classes (Top to bottom: Adenocarcinoma, Large Cell Carcinoma, Normal, Squamous Cell Carcinoma)

A close analysis of Figure 2 shows the distinct characteristics of each class. The Adenocarcinoma classes have chest walls that are swollen into odd shapes, while the pulmonary columns of some scans are attached to the chest walls by a lung mass. Similarly, the Large Cell Carcinoma images depict pulmonary columns swollen to a larger size than typically

seen in a lung. Additionally, the lung nodules are larger than those in a normal lung, approaching the point where all such tissues are lung masses rather than nodules. Further, the chest walls in this category are more distorted than those in the other classes. The normal class has chest walls and pulmonary columns which are smaller in size than the other classes. Additionally, there are fewer lung nodules, with some scans showing no visible nodules at all. Furthermore, the Squamous Cell Carcinoma scans have chest walls that are swollen and deformed, while the area surrounding the pulmonary columns is occupied by a giant mass in some of the scans. In others, the chest walls remain distorted, but the pulmonary columns are more prominent.

2.2.2 *Average Images*

The average image for each class was created using the Python mean function, which averaged the scans from each category to generate an average image. These images show the overall characteristics of each class, which can be compared with each other to understand them further.

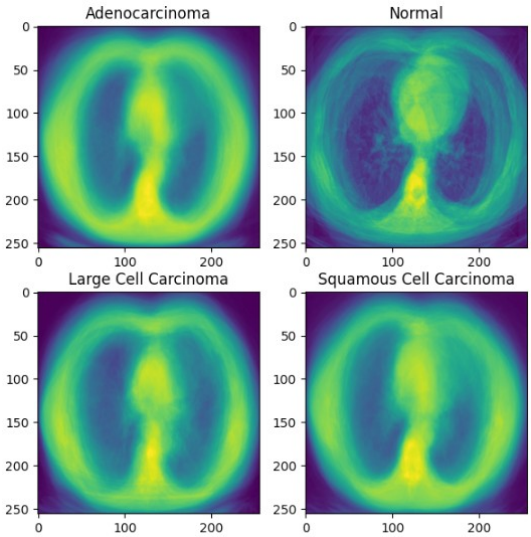

Figure 3. A visual representation of average images for each of the four classes

Figure 3 shows several significant distinctions and similarities between the classes. The normal lung scans have a less consistent shape, with a higher variation in tissue placement. In contrast, the three cancer classes have defined chest walls and pulmonary columns. Though distinctions between each separate class are less obvious, some characteristics can be pointed out. The Squamous Cell Carcinoma scan has a higher concentration of lung masses on the top right side than the other three classes. Additionally, the Large Cell Carcinoma image seems highly saturated in the center of the pulmonary column. The Adenocarcinoma scan is the most concentrated at the bottom of the pulmonary column, like the Squamous Cell Carcinoma. These highly saturated areas may suggest each class's tumor location or shared tissue structures.

2.2.3 *Average Image Pixel Distribution*

Using Matplotlib (8), the flattened pixel distribution of the average images can be visualized as histograms. These charts compare the shape and structure of each class, providing a

unique perspective on the characteristics of the classes.

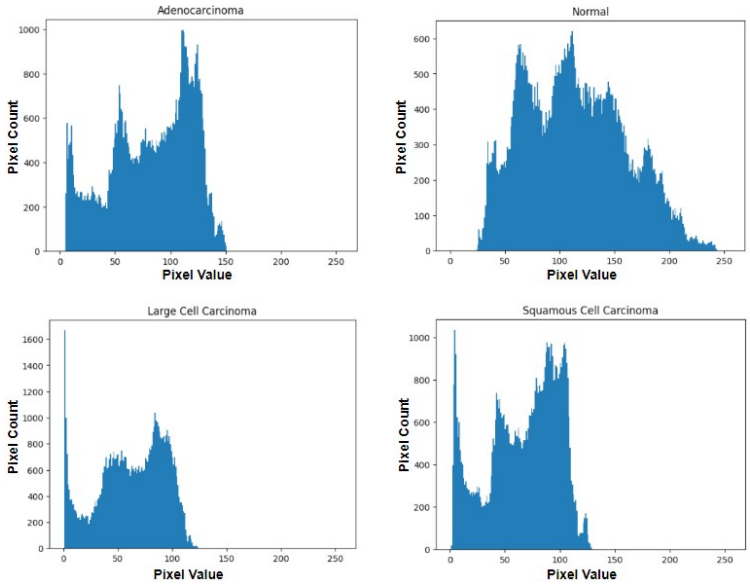

Figure 4. A visual representation of average image pixel distribution for each of the four classes

Figure 4 displays the concentration of tissues in certain areas of the average images, revealing characteristics of the classes from a new perspective. The three cancer classes have pixel count values on the y-axis ranging from 0 to >1000, occurring between the pixel values of 0 to 150. However, the normal class has a smaller range, from ~ 0 to 600 pixels, spread out throughout the entire x-axis. This means that the malignant average images have greater concentrations of lighter pixels in certain areas, while that is a greater variation of pixel tone and placement in the benign average image. Figure 3 confirms this analysis, as the cancerous average images appear more concentrated, forming a definite overall shape, while the normal image is more varied. Additionally, the graphs are skewed slightly to the right, which is more apparent in the normal class. This represents the negative space around the lungs in the scans, which can be seen in all the classes. These observations complement the attributes seen in Figure 3, where certain areas are denser than others.

2.3 Modeling

Several models were used to attain the highest accuracy and efficiency. These models included architecture and methods such as CNNs, data augmentation, and transfer learning. Specifically, sections 2.3.1 through 2.3.3 discuss the models that served as the building blocks for the main objects of comparison – the transfer learning models. Thus, it is important to note that the main models being compared are the transfer learning models, as they have been trained using ImageNet (9), while the other models have not. This section will explore the details and findings of each model and compare their results, focusing on the performance of the transfer learning models.

2.3.1 *Vanilla Model*

The vanilla model was the first attempt at modeling. It used a simple neural network architecture with a 2D convolutional layer, two dense layers, a dropout layer, and a max pooling layer. After training, the model ended with a training accuracy of 99.84% and a test accuracy of 62.5%.

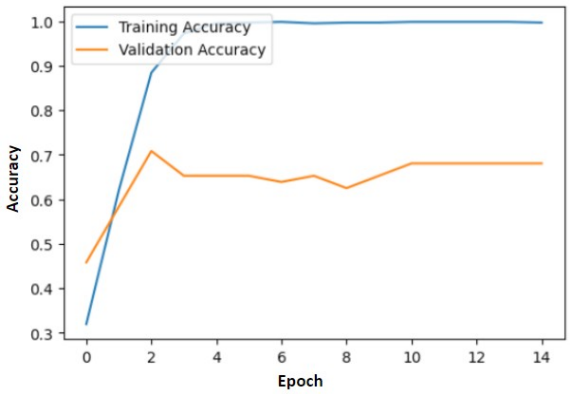

Figure 5. Accuracy Versus Epoch Plot for the Vanilla Model

However, as is seen in Figure 5, the basic architecture of this model caused it to be prone to overfitting, which occurs when the model performs at a high training accuracy, but a low validation accuracy due to the new data introduced in the validation set. Specifically, this model had a high training and low test accuracy, meaning the 99.84% accuracy was inaccurate since the test accuracy itself was only 62.5%. Thus, the vanilla model is not a desirable choice in the real world due to its low accuracy.

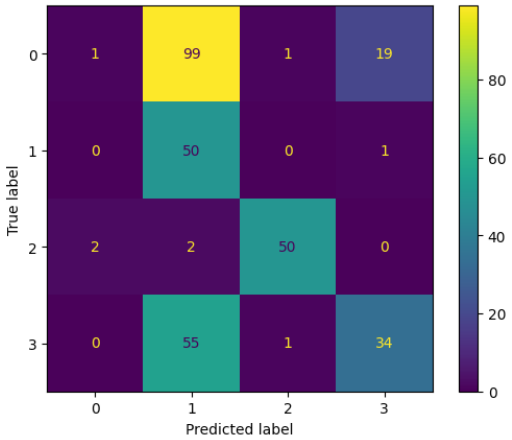

Figure 6. Model Confusion Matrix for the Vanilla Model

Figure 6 portrays the vanilla model's low accuracy. There was a total of 135 correctly predicted classes out of the 315 total images, which are shown by the results along the diagonal reaching from the top left of the matrix to the bottom right. The model most frequently misclassified the Adenocarcinoma images as Large Cell Carcinoma. This means the model requires more data for these two classes for it to properly classify them.

2.3.2 *Model with Data Augmentation*

The model with data augmentation was built using the vanilla model as its foundation, with the addition of augmented images to reduce overfitting. Specifically, the augmentations used included random flip, random rotation, and rescaling. After training, the model ended with a training accuracy of 69.98% and a test accuracy of 63.89%.

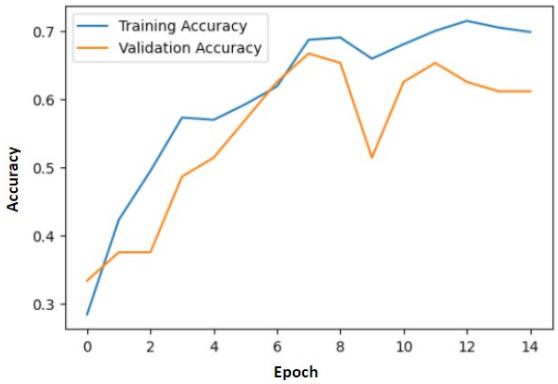

Figure 7. Accuracy Versus Epoch Plot for the Model with Data Augmentation

As is shown in Figure 7, the training and test accuracies are close to each other, meaning there is no overfitting. However, despite the decrease in fitting rate, the model still performed at a low accuracy, disfavoring it, among others.

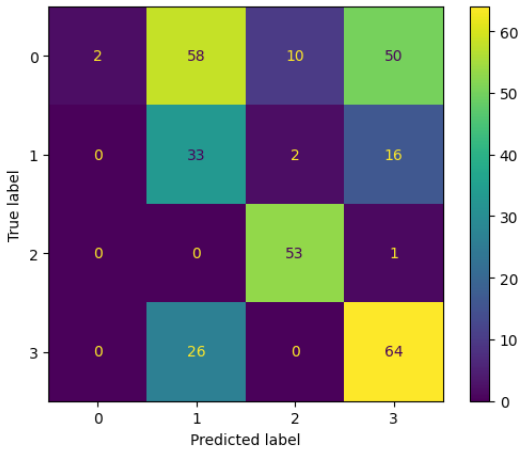

Figure 8. Model Confusion Matrix for the Model with Data Augmentation

Figure 8 depicts the accuracy for each class. The diagonal axis shows that 152 images were correctly predicted out of the 315 total scans the model analyzed. The model seems suitable for classifying scans in normal class. However, it misclassified the 3 cancer classes frequently, particularly for the Adenocarcinoma class which was entirely misclassified.

2.3.3 *Model with Dynamic Learning Rate and Early Stopping*

In another attempt to reduce overfitting and improve accuracy, we introduced a dynamic

learning rate and early stopping. The learning rate is the crucial hyperparameter used to train deep convolutional neural networks. The dynamic learning rate reduces the speed at which the model learns when the accuracy hits a plateau. The performance of the model thus improves and attains comparatively high accuracy with fewer iterations (10). We also introduced early stopping, a method that allows for the automatic stopping of training once the model performance stops improving on the validation dataset (11). This model utilized the vanilla model as a foundation and used these two callbacks to work more accurately and efficiently. After being trained, the model ended with a training accuracy of 92.17% and a test accuracy of 63.89%. In this case, the model again overfitted as seen in the high training accuracy and low test accuracy.

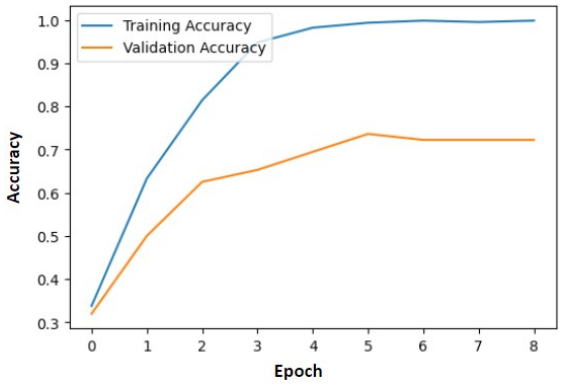

Figure 9. Accuracy Versus Epoch Plot for the Model with Dynamic Learning Rate and Early Stopping

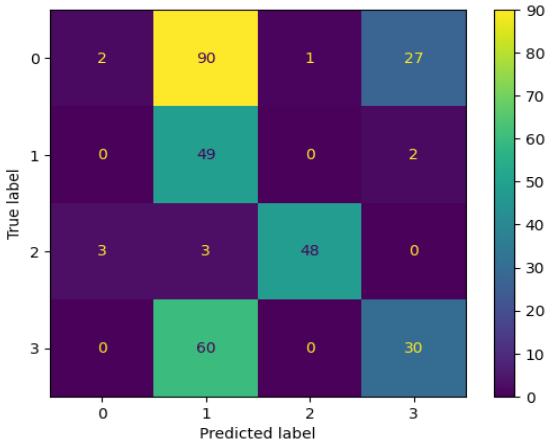

Figure 10. Model Confusion Matrix for the Model with Dynamic Learning Rate and Early Stopping

Figure 10 depicts the accuracy distribution for each class. The diagonal axis shows that 129 images were correctly predicted out of the 315 total scans the model analyzed. The model is suitable for classifying images from the normal class compared to the cancer classes. This especially applies to the Squamous Cell Carcinoma class, classified as Large Cell Carcinoma 60 times.

2.3.4 *Transfer Learning*

Transfer learning is a technique in machine learning and deep learning that involves utilizing pre-existing knowledge gained from one task to improve the performance of another related task. Transfer learning is particularly useful when there is limited labeled data available for the target task. It can be applied in two ways: as a baseline algorithm, used to train the image dataset and evaluate performance; and as a feature extractor, where features are extracted from image datasets and used with machine learning or deep learning algorithms to assess the performance. The three transfer learning models tested included VGG16 (12), VGG19 (12), and ResNet50 (13). In each case, the model was pre-trained using Image Net (9), which is a large database of annotated images used for visual object recognition research.

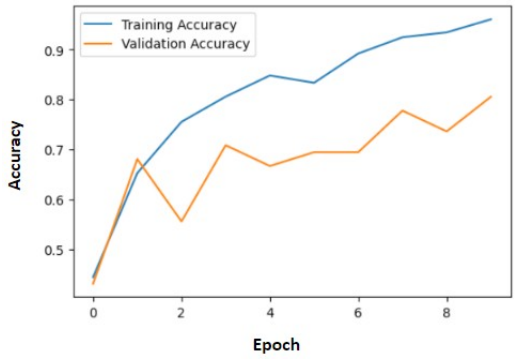

Figure 11. Accuracy Versus Epoch Plot for the VGG16 Model

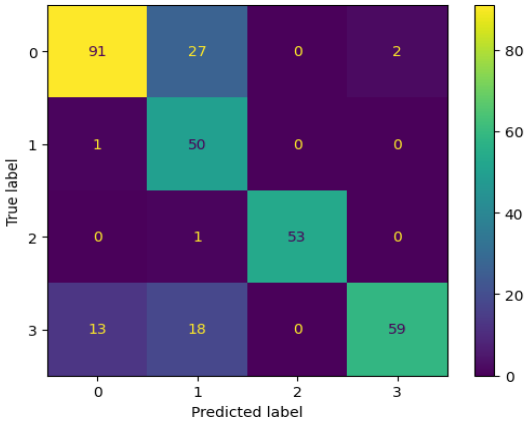

Figure 12. Model Confusion Matrix for the VGG16 Model

Figure 12 depicts the accuracy distribution for each class. The diagonal axis shows that 253 images were correctly predicted out of the 315 total scans the model analyzed. A major improvement in this model was the increased accuracy for classifying Adenocarcinoma images. However, the model misclassified the Squamous Cell Carcinoma class the most, often predicting it as Adenocarcinoma or Large Cell Carcinoma. The VGG16 model is a 16-layer model with pre-trained weights. Stacked on top of this model are three dense layers, two dropout layers, and a flattening layer used for fine-tuning. A limitation posed by this architecture was the training period. Compared to the other models, the VGG16 took an immense amount of time to train. Despite the lengthy training time,

the model ended with a training accuracy of 96.08% and a test accuracy of 80.56%.

The VGG19 model is a 19-layer model with pre-trained weights. Stacked on top of this model are three dense layers, two dropout layers, and a flattening layer used for fine-tuning. Like the VGG16 architecture, this model took a lengthy time to train. At the end of training, the model ended with a training accuracy of 90.86% and a test accuracy of 73.61%.

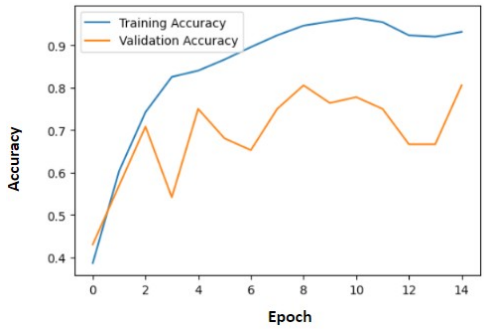

Figure 13. Accuracy Versus Epoch Plot for the VGG19 Model

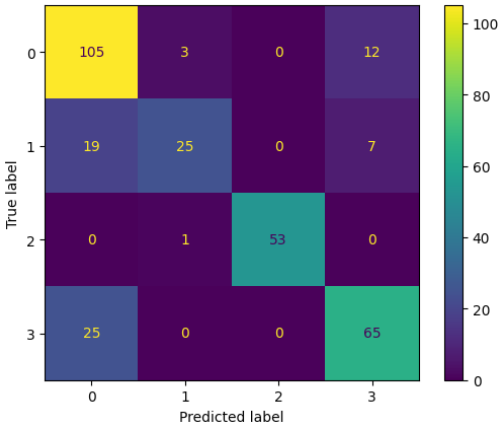

Figure 14. Model Confusion Matrix for the VGG19 Model

Figure 14 depicts the accuracy distribution for each class. The diagonal axis shows that 248 images were correctly predicted out of the 315 total scans the model analyzed. The model is best suited for classifying Adenocarcinoma images, with 105 correct predictions for this class. However, the Large Cell Carcinoma and Squamous Cell Carcinoma classes were the most misclassified, indicating the necessity for more data from these two classes.

The ResNet50 model is a 50-layer model with pre-trained weights. Stacked on top of this model are three dense layers, two dropout layers, and a flattening layer for fine-tuning. In addition to these layers, the model used the data augmentation mentioned in Section 2.1.1. The ResNet50 architecture resulted in the most successful performance, with a training accuracy of 99.18% and a test accuracy of 88.89%.

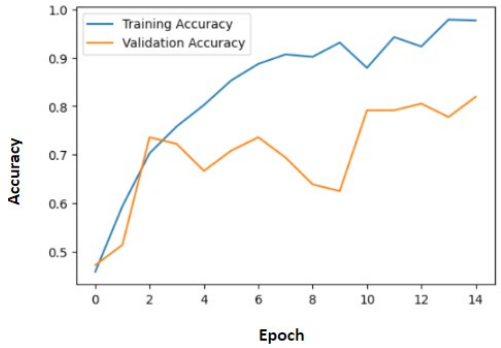

Figure 15. Accuracy Versus Epoch Plot for the ResNet50 Model

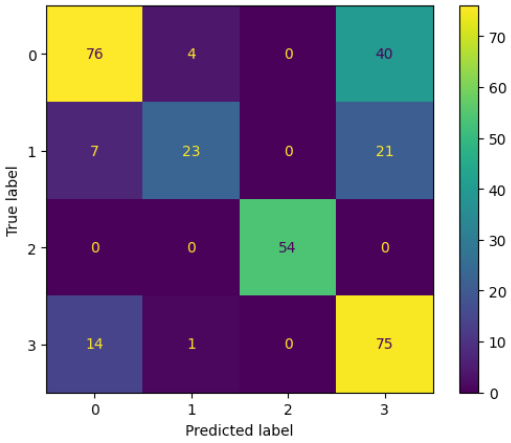

Figure 16. Model Confusion Matrix for the ResNet50 Model

Figure 16 depicts the accuracy distribution for each class. The diagonal axis shows that 228 images were correctly predicted out of the 315 total scans. The model is best suited for classifying normal images. and has the most balanced results compared to the other models due to fewer false positive and negative predictions.

3 Results and Discussion

3.1 Results

Table 1: Model Performance

Model	Training Accuracy	Test Accuracy
Vanilla Model	99.84	62.50
Model with Data Augmentation	69.98	63.89
Model with Dynamic Learning and Early Stopping	92.17	63.89
VGG16	96.08	80.56
VGG19	90.86	73.61
ResNet50	99.81	88.89

Table 2: Model Precision by Class

Model	Class 1	Class 2	Class 3	Class 4
Vanilla Model	0.33	0.24	0.96	0.63
Data Augmentation	1.00	0.28	0.82	0.49
Dynamic Learning and Early Stopping	0.40	0.24	0.98	0.51
VGG16	0.87	0.52	1.00	0.97
VGG19	0.88	0.65	1.00	0.82
ResNet50	0.78	0.82	1.00	0.83

Table 3: Model Recall by Class

Model	Class 1	Class 2	Class 3	Class 4
Vanilla Model	0.01	0.98	0.93	0.38
Data Augmentation	0.02	0.65	0.98	0.71
Dynamic Learning and Early Stopping	0.02	0.96	0.89	0.33
VGG16	0.76	0.98	0.98	0.66
VGG19	0.88	0.49	0.98	0.72
ResNet50	0.87	0.45	1.00	0.83

3.1.1 *Model Classification Report*

This section focuses on analyzing the precision and recall for each model. Precision is a metric that measures how often a machine learning model correctly predicts the positive class (14). A higher precision means a lower number of false positives, correlating to the number of falsely detected lung cancers in real life. On the other hand, recall is a metric that measures how often a machine learning model correctly identifies positive instances from all the actual positive samples in the dataset (14). A higher recall indicates a lower amount of false negative predictions. In medical settings, failing to diagnose cancer is more detrimental than receiving erroneous treatment, thus resulting in the prioritization of recall over precision. Table 1 quantifies each model's performance, with the ResNet50 model with an 88.89% testing accuracy. Table 2 portrays the class precision for each model tested during the project. All three transfer learning models demonstrated 100% precision for benign images. However, the ResNet50 model showed the most balanced precisions of the three models. Table 3 shows the recall rates for each model. Like Table 2, all three transfer learning models demonstrated high recalls for benign images, with ResNet50 portraying the most balanced results. In conclusion, all three tables showed that the ResNet50 model performed the best, with the highest accuracy and most balanced precision and recall.

3.2 Project Limitations

A common challenge faced by all medical research is the paucity of data. In the case of this project, the CT scans are derived from real patients who consented to publishing these images. However, this information is often limited, making it difficult to train a model due to the lack of data. The scans used in this project were compiled from several different sources, which also makes it hard to verify their

credibility. In the future, it would be beneficial to use data from a single reputable source to deliver the highest performance. Additionally, this data under-represented and over-represented certain classes. As stated in Section 2.1.3, the slice count for Large Cell Carcinoma was much lower than the rest, causing imbalances in the model performance. Regarding the models, certain methods such as L1 and L2 regularization were not implemented in the architecture, as the data used was limited. Additionally, during the development of the model, the prevalence of false-positives and false-negatives was a major criterion, as a false-negative is far more dangerous than a false-positive classification, as was discussed in Section 3.1.1. This model has more false-positives than false-negatives, thus in the case that a patient does not have cancer, the model will output a false positive and a medical professional will be able to determine the patient's true condition using other methods. However, the model remains limited to classifying the three cancer types.

4 Conclusion

A series of machine learning models were developed to detect the presence and type of lung cancer based on CT scans. The highest-performing model was found to be a fine-tuned ResNet50, predicting the presence and cancer type with a test accuracy of 88.89%. This machine learning model can serve as an aid for doctors to correctly detect and dignose lung cancer.

5 References

1. Lung Cancer Statistics | LUNGevity Foundation. (n.d.). Www.lungevity.org. https://www.lungevity.org/for-supporters-advocates/lung-cancer-awareness/lung-cancer-statistics

2. Nazir, I., Haq, I. ul, AlQahtani, S. A., Jadoon, M. M., Dahshan, M. (2023). Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion. Journal of Sensors, 2023, e6683438. https://doi.org/10.1155/2023/6683438

3. MathWorks. (2023). What is a Convolutional Neural Network? Www.mathworks.com. https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html

4. Adenocarcinoma - types and treatment options. (2019, May 31). Cancer Treatment Centers of America. https://www.cancercenter.com/adenocarcinoma

5. Large cell carcinoma: Stages, treatment, prognosis, and more. (2021, March 4). Www.medicalnewstoday.com. https://www.medicalnewstoday.com/articles/large-cell-carcinoma

6. Group, L. C. (2023, March 23). Squamous Cell Carcinoma Lung Cancer | Causes, Diagnosis, & Treatment. Lung Cancer Group. https://www.lungcancergroup.com/lung-cancer/non-small-cell-lung-cancer/squamous-cell-carcinoma/

7. Hany, Mohamed. (2020). Chest CT-Scan images Dataset [Data set]. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images

8. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/mcse.2007.55

9. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2009.5206848

10. Johny, A., Madhusoodanan, K. N. (2021). Dynamic Learning Rate in Deep CNN Model for Metastasis Detection and Classification of Histopathology Images. Computational and Mathematical Methods in Medicine, 2021, e5557168. https://doi.org/10.1155/2021/5557168

11. Vijay, U. (2020, October 13). Early Stopping to avoid overfitting in neural network- Keras. Medium. https://medium.com/zero-equals-false/early-stopping-to-avoid-overfitting-in-neural-network-keras-b68c96ed05d9

12. Simonyan, K., Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv.org. https://arxiv.org/abs/1409.1556v6

13. He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep Residual Learning for Image Recognition. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1512.03385

14. Accuracy vs. precision vs. recall in machine learning: what's the difference? www.evidentlyai.com. https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall