Lung and Colon Cancer Detection Using Histopathological Images

EE 628 WS - Deep Learning

Final Project Report

Instructor: Rensheng Wang

Author:

Hari Kiran Vaddi Gangaraju - CWID 20021466

1. Introduction

Lung and colon cancers remain the leading causes of cancer-related deaths globally (American Cancer Society, 2023). Accurate and early detection significantly increases survival rates, but traditional diagnostic methods depend on manual examination by pathologists, which introduces delays and potential errors. This project aims to develop a deep learning model to automate the detection of lung and colon cancer from histopathological images. By leveraging the power of Convolutional Neural Networks (CNNs), we provide a tool that enhances diagnostic efficiency and accuracy.

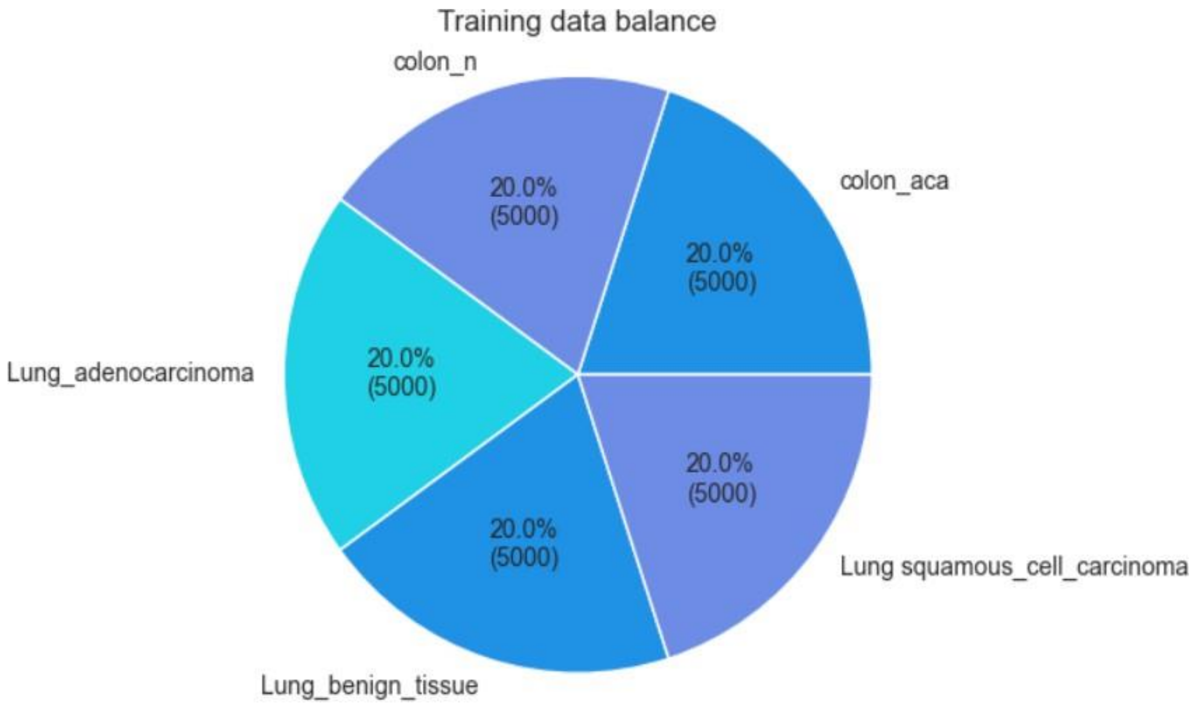
2. Data Collection and Preprocessing

2.1 Data Source:

We used a publicly available dataset from Kaggle that complies with HIPAA regulations. The original dataset contained 750 histopathological images divided into five categories:

- Lung benign tissue
- Lung adenocarcinoma
- Lung squamous cell carcinoma
- Colon adenocarcinoma
- Colon benign tissue

To increase the dataset size and improve the model's ability to generalize, we applied data augmentation. This process expanded the dataset to 25,000 images, with all images retaining their original resolution of 768x768 pixels. Distribution of Dataset can be seen here



2.2 Data Preprocessing

We carefully prepared the data before feeding it into the model. At first, we checked the class distribution and confirmed that all five categories were well-represented. We split the dataset into three subsets 80% for training, 10% for validation,and 10% for testing. We normalized the pixel values of the images to a range of 0 to 1. This step improved both training efficiency and model performance.

3. Model Development

3.1 Model Architecture:

We built a CNN model to classify the histopathological images. The architecture includes the following components:

- **Input Layer**: Accepts 768x768 pixel images.

- **Convolutional Layers**: Extract spatial features from the images using filters.
- **Batch Normalization**: Normalizes intermediate outputs to stabilize the training process.
- **Max Pooling Layers**: Reduces the size of the feature maps while retaining important features, helping to prevent overfitting.
- **Dense Layers**: Captures complex relationships between features to improve classification accuracy.
- **Dropout Layers**: Adds regularization by randomly deactivating neurons during training.
- **Output Layer**: Uses a softmax activation function to classify images into one of the five categories.

This architecture includes:

- Convolutional Blocks: To extract spatial features at various levels.
- Batch Normalization: To stabilize and accelerate training.
- Max Pooling: To reduce feature dimensions and prevent overfitting.
- Dense Layers with Dropout: For feature extraction and regularization.
- Softmax Output Layer: For multi-class classification.

3.2 Training Process:

We trained the model with the following settings:

- Loss Function: Categorical cross-entropy, which works well for multi-class classification.
- Optimizer: Adam, combined a learning rate scheduler to adjust the learning rate dynamically.

- Epochs: We trained the model until the validation loss stabilized, ensuring the model didn’t overfit or underfit.

We monitored the validation metrics throughout the training process to confirm the model was learning effectively.

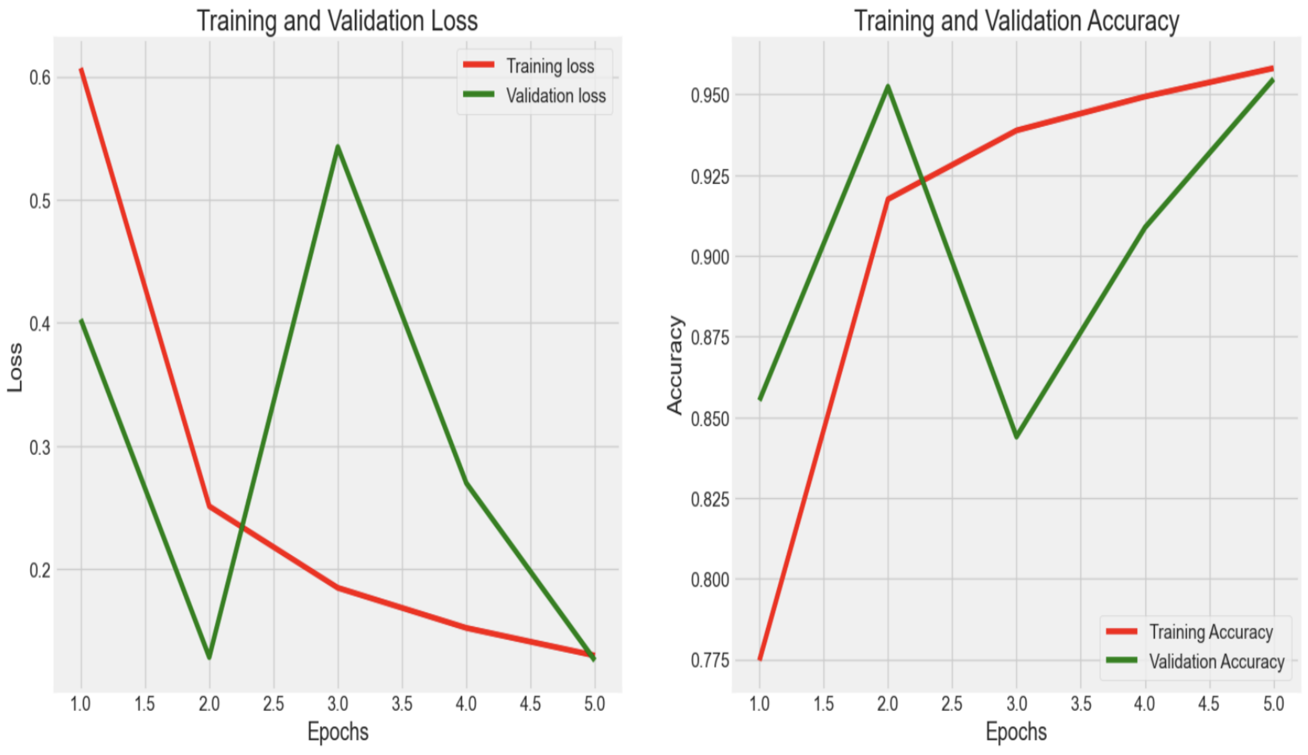
4. Results

4.1 Performance Metrics:

We evaluated the model’s performance using several key metrics:

1) Accuracy: Measures the percentage of correct predictions.
2) Loss: Represents the error between the model’s predictions and actual labels.
3) Confusion Matrix: Provides a detailed breakdown of correct and incorrect classifications for each class.

The graphs below illustrate the training and validation accuracy and loss over epochs:



4.2 Training, Validation, and Test Results

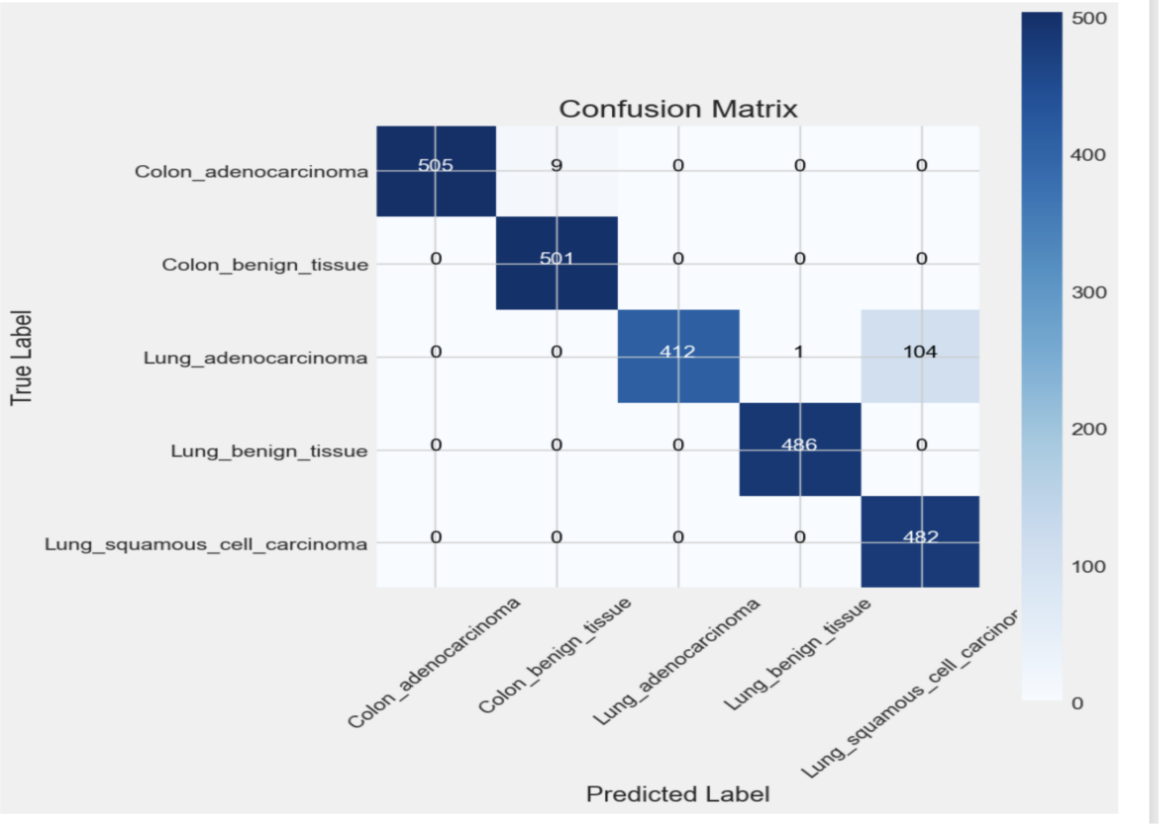
Our CNN achieved the following results:

- Training Accuracy: 95.96%
- Validation Accuracy: 95.48%
- Test Accuracy: 95.44%

The similar performance across training, validation, and test datasets shows that the model generalizes well to unseen data. Loss values remained low across all datasets, confirming the model avoided both overfitting and underfitting.

4.3 Confusion Matrix Analysis

The confusion matrix showed strong classification performance for most classes. However, the model occasionally misclassified closely related subtypes, such as "colon adenocarcinoma" and "colon benign tissue." These errors suggest areas where the model could improve further. The confusion matrix below summarizes the classification performance, highlighting areas of strength and misclassification:



Conclusion

This project successfully built an automated cancer detection system using CNNs. The model achieved high accuracy across all datasets, demonstrating its ability to classify histopathological images effectively.

By reducing the dependency on manual diagnosis, this system can speed up the diagnostic process and support healthcare professionals in making more accurate decisions. While the model performed well, further improvements could enhance its ability to distinguish between similar subtypes. For example, future work could include testing advanced architectures or applying domain-specific preprocessing techniques.

In conclusion, this project highlights the potential of deep learning to address critical challenges in medical image analysis. By integrating this system into clinical workflows, we can improve early cancer detection and patient outcomes.