Project Title

Lung Cancer Classification Using Histopathology Images

Darya Ardan

Objective:

The main objective of this project was to develop and evaluate classification models for identifying lung cancer types and tissues using histopathology images from the LC25000 dataset. The data is taken from Kaggle.com:

"https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images"

Description:

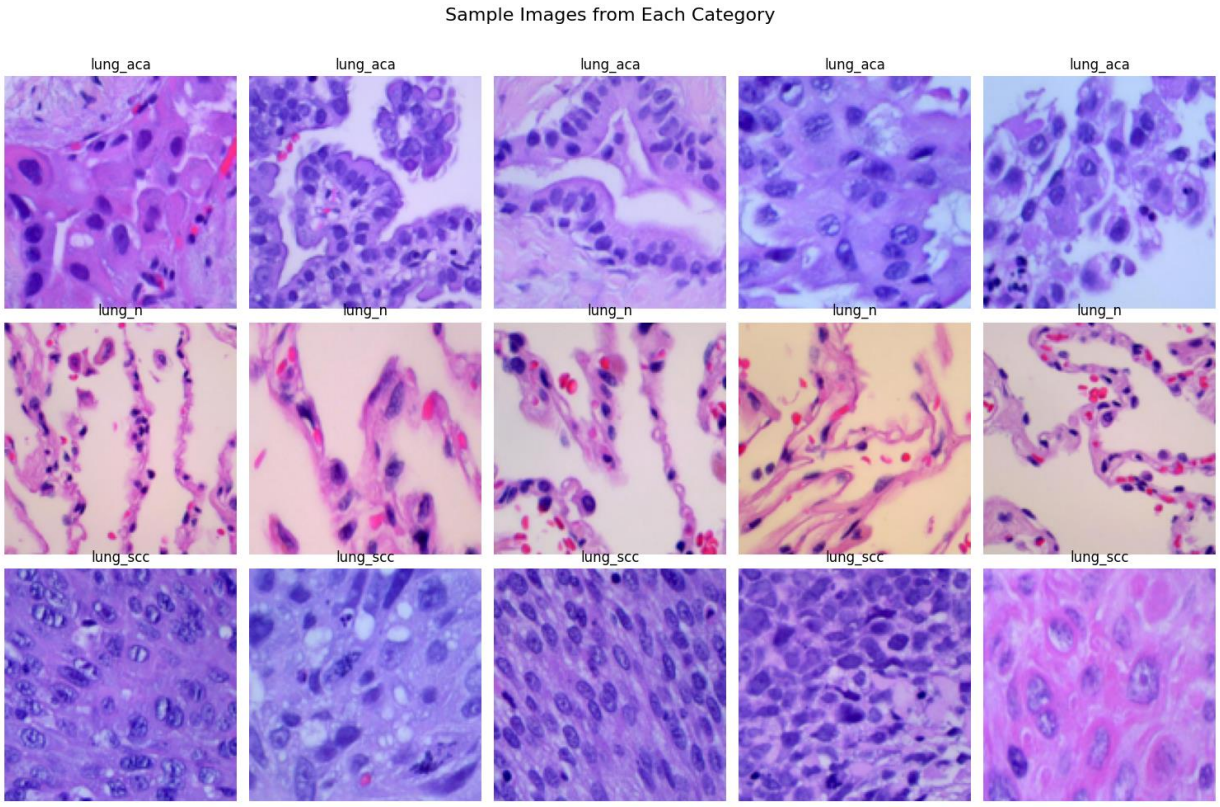
This project focuses on the classification of lung cancer using histopathology images. The LC25000 dataset, which contains images from lung and colon tissue sections, was used, specifically focusing on the lung images. The goal was to build robust models that can accurately classify different types of lung cancer and tissues.

Technologies Used:

- **Programming Language**: Python
- **Libraries**: PyTorch, OpenCV, NumPy, PIL (Python Imaging Library)
- **Models**: Custom CNN models, Pre-trained ResNet50
- **Tools**: Python IDE, Visual Studio Code

Data:

The LC25000 dataset comprises histopathology images of lung and colon tissues. For this project, only the lung tissue images were utilized. These images are suitable for building classification models to distinguish between different types of lung cancer and tissues. The codes are available in Histo_data.py



Pre-processing:

Pre-processing of histopathology images is crucial to enhance image quality and extract relevant features. The following steps were undertaken:

1. **Histogram Equalization**: This technique was used to improve the contrast of the images by distributing the intensity values more evenly.
2. **Contrast Limited Adaptive Histogram Equalization (CLAHE)**: CLAHE was applied to further enhance the local contrast of the images, making it easier to distinguish different tissue structures.

3. **Random Morphological Transformations**: Random dilation and erosion operations were applied to augment the dataset and improve the robustness of the models.
4. **Image Transformations**: Images were resized, randomly rotated, cropped, and flipped horizontally and vertically to create a diverse set of training samples. The images were then normalized to standardize the input for the models.

The codes are available in data_loader.py

Model Development:

Three different models were developed and evaluated:

1. **Custom CNN Models**: Two different custom Convolutional Neural Network (CNN) models, named SimpleCNN and CNN2, were designed specifically for this task. These models were trained and evaluated on the processed dataset. The codes are available in modelCNN.py
2. **Adjusted Pre-trained ResNet50**: An adjusted pre-trained ResNet50 model was fine-tuned on the dataset. This model leverages the power of deep pre-trained features and transfer learning.

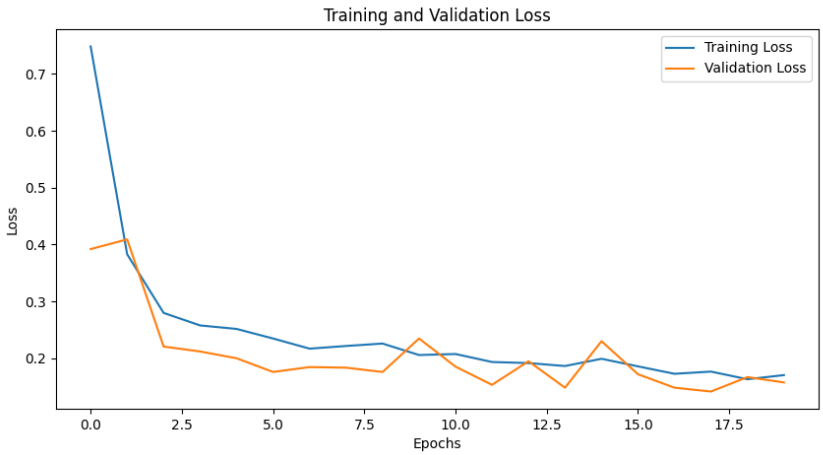
Training and Evaluation

The models were trained on train dataset and evaluated their performance on the validation dataset the pre-processed dataset. The following metrics were used for evaluation:

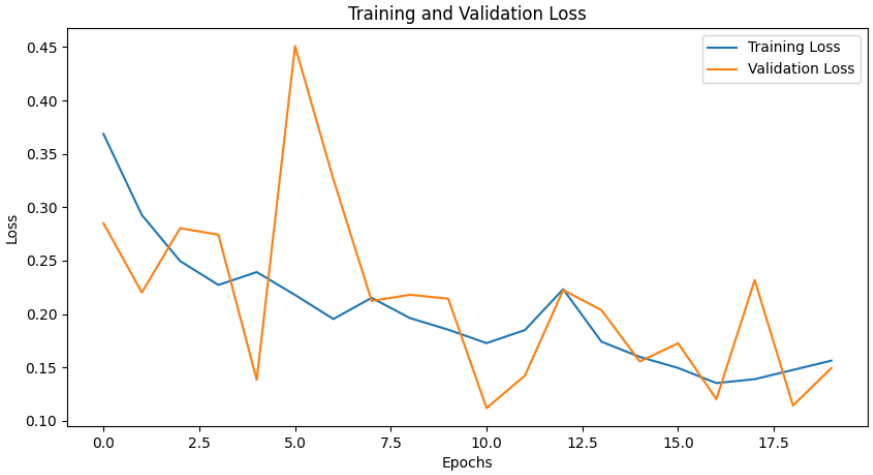
- **Accuracy**: The accuracy of the custom CNN models was approximately 93%, while the ResNet50 model achieved an accuracy of around 94%.
- **Loss**: The loss of each model was plotted to visualize the training process and compare their performance on the test set.

All three models trained for 20 epochs and the batch size of 64. Below you can see the training and validation loss progress of each model:

SimpleCNN:



CNN2:

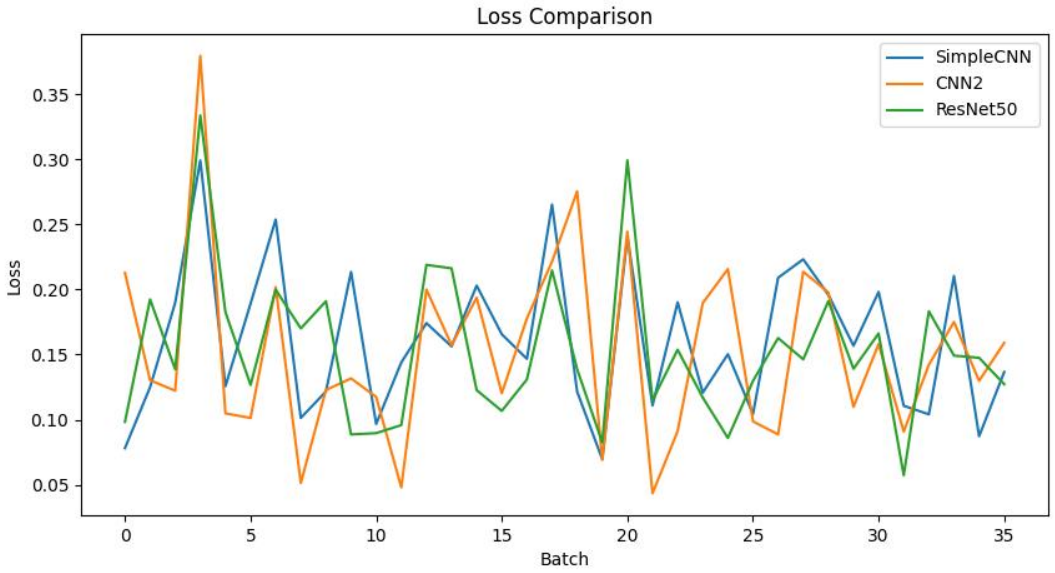


Adjusted Resnet50:

Results:

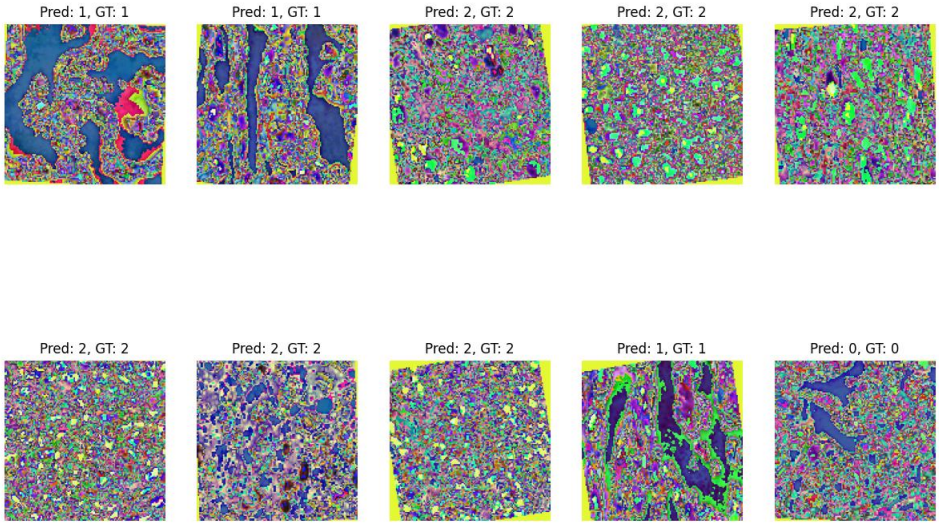
- **Custom CNN Models**: Achieved around 93% accuracy.
- **Pre-trained ResNet50**: Achieved approximately 94% accuracy.

The models were also evaluated on unseen test data to ensure their generalizability. Similarly, custom CNN models got around 93% accuracy and adjusted Resnet50 got 94% accuracy. The plot below shows the comparison of the loss of three models.

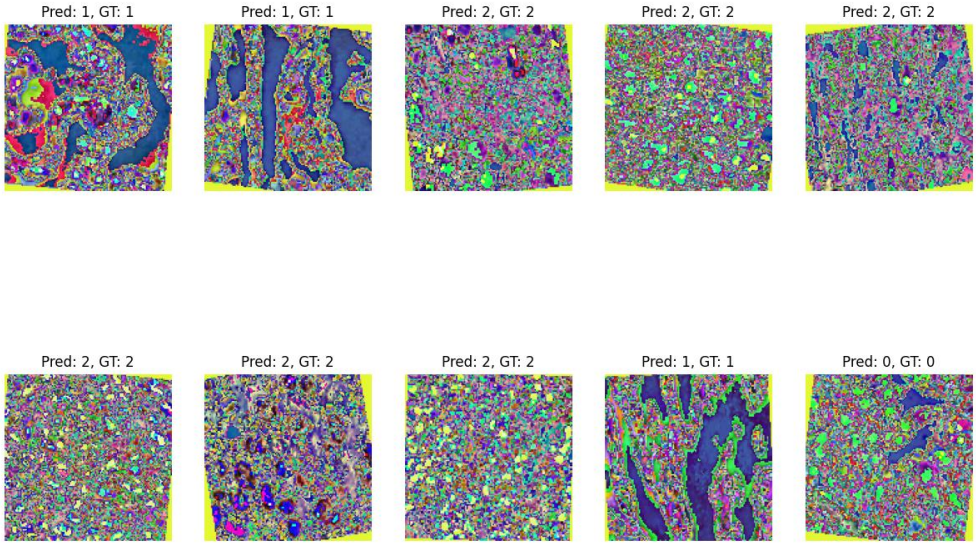


Sample predictions with their ground truth labels were displayed to illustrate the performance of the models.

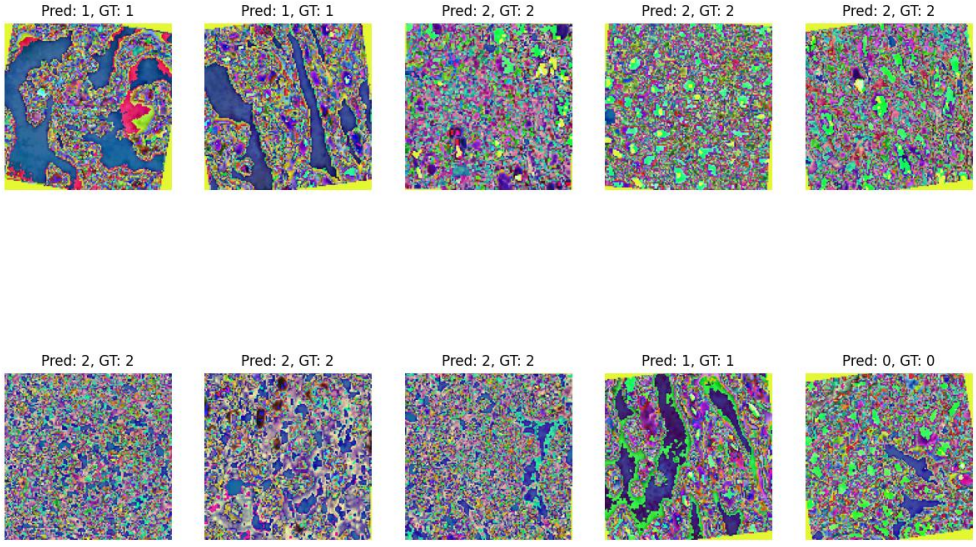
SimpleCNN:



CNN2:



Adjusted Resnet50:



Conclusion

The project successfully demonstrated the ability to classify lung cancer histopathology images with high accuracy using both custom-designed CNN models and a pre-trained ResNet50 model. The pre-processing steps, including histogram equalization and CLAHE, played a significant role in enhancing image quality and improving model performance. Future work could involve exploring other pre-trained models and additional data augmentation techniques to further improve classification accuracy.