TABLE OF CONTENTS

Abstract

This research project presents a deep learning-based system for the early detection and classification of lung cancer using histopathological images. Histopathological images are high-resolution microscopic images of tissue samples, commonly used by pathologists to analyze morphological changes and diagnose diseases such as cancer. According to the World Health Organization (WHO), lung cancer accounts for approximately 2.2 million new cases and 1.8 million deaths annually, making it the leading cause of cancer mortality worldwide.

In this study, we implement a fine-tuned InceptionV3 convolutional neural network to distinguish between three classes: Normal Lung Tissue, Squamous Cell Carcinoma, and Adenocarcinoma. The initial dataset consisted of only 250 labeled images, which was augmented through transformations such as rotation, flipping, and zooming to produce a dataset of approximately 5000 images. The dataset was preprocessed, stratified, and fed into a modified InceptionV3 architecture with the top layers unfrozen for fine-tuning. The system was evaluated using accuracy/loss plots, confusion matrix, and classification metrics, achieving exceptional performance, suggesting its practical applicability for diagnostic support.

Introduction

Lung cancer is the most commonly diagnosed cancer and the leading cause of cancer-related deaths globally. It is broadly categorized into two major types: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), with NSCLC accounting for about 85% of all cases. Within NSCLC, adenocarcinoma and squamous cell carcinoma are the most prevalent subtypes. Early and accurate detection of these subtypes significantly improves the chances of successful treatment.

Histopathological examination, which involves the microscopic analysis of stained tissue samples, remains the gold standard for lung cancer diagnosis. However, manual interpretation of these high-resolution images is time-consuming and requires a high level of expertise. The subjectivity involved may lead to inconsistencies and diagnostic delays.

With the growing integration of artificial intelligence in healthcare, deep learning has emerged as a powerful tool in medical image analysis. Convolutional neural networks (CNNs), in particular, have shown tremendous promise in extracting spatial hierarchies of patterns. InceptionV3 is a widely used CNN model known for its architectural depth and computational efficiency.

This project explores the potential of a fine-tuned InceptionV3 model for accurately classifying histopathological images of lung tissue into three categories: Normal, Squamous Cell Carcinoma, and Adenocarcinoma. The methodology includes data augmentation, model construction, training, and evaluation, aiming to create a robust diagnostic support tool.

Problem Statement

Lung cancer is one of the most aggressive and fatal malignancies worldwide, with its prognosis heavily dependent on early detection. Traditional diagnostic techniques such as radiology and histopathological evaluation require significant expertise and time, and are often subject to inter-observer variability. Histopathology, in particular, involves microscopic examination of stained tissue samples and is considered the gold standard for confirming cancer. However, this manual process is inherently slow, labor-intensive, and susceptible to human error due to fatigue and subjectivity.

With the increasing volume of clinical data and a shortage of expert pathologists, there is a critical need for reliable, automated tools to assist in diagnostic workflows. The aim of this project is to develop a deep learning-based solution to classify lung tissue histopathological images into three major categories: Normal, Squamous Cell Carcinoma, and Adenocarcinoma. This automated approach can potentially reduce diagnostic time, enhance consistency, and support medical professionals by highlighting areas of concern with high accuracy and reliability. By fine-tuning a robust convolutional neural network like InceptionV3 on domain-specific histopathological data, this system seeks to bring scalable intelligence into digital pathology applications.

Literature Survey

Recent advancements in deep learning have significantly enhanced the classification of lung cancer using histopathological images. Notable studies include:

- **Deep Convolutional Neural Networks for Lung Cancer Classification**: A study published in 2023 utilized deep convolutional neural networks (DCNNs) to classify lung cancer subtypes—adenocarcinoma, squamous cell carcinoma, and small cell carcinoma—achieving an accuracy of 71.1%. This research underscores the potential of DCNNs in automating lung cancer diagnosis. Wiley Online Library

- **Explainable AI for Lung Cancer Detection**: A recent publication in *Scientific Reports* introduced an explainable AI model using a custom CNN on CT images for early and accurate lung cancer detection. The study emphasizes the importance of interpretability in AI models for clinical applications. Nature

- **ELW-CNN for Automated Detection**: An extremely lightweight convolutional neural network (ELW-CNN) was proposed for the automated detection of lung and colon cancers through histological analysis. The model focuses on reducing computational complexity while maintaining high accuracy. PMC

- **LungHist700 Dataset Introduction**: A 2024 study presented the LungHist700 dataset, comprising 691 high-resolution histopathological lung images, including adenocarcinomas, squamous cell carcinomas, and normal tissues. This dataset serves as a valuable resource for training and evaluating deep learning models. Nature

- **AI-Powered Lung Cancer Detection Using VGG16**: Research published in *Information* demonstrated the effectiveness of the pre-trained VGG16 model in detecting lung cancer from histopathological images, highlighting its potential in enhancing diagnostic precision. MDPI

- **Harnessing Transformers for Lung Cancer Detection**: A 2023 study explored the application of transformer-based models in lung cancer image detection, achieving an accuracy of 94.71% for histopathological lung cancer classification, indicating a promising direction for future research. arXiv

- **Deep Learning for Lung Cancer Diagnosis and Prognosis**: A systematic review in *Cancers* provided an overview of current advances in deep learning-based methods for lung cancer diagnosis, subtyping, prognosis prediction, and mutational status characterization using histological and cytological images. MDPI

- **Classification and Mutation Prediction Using Inception V3**: A study utilized the Inception V3 architecture to classify non-small cell lung cancer and predict mutations from histopathology slides, demonstrating the model's capability in both classification and mutation prediction tasks. PMC

Hardware and Software Requirements

Hardware Requirements

To successfully implement and run the lung cancer detection model, a system with moderate to high computational capabilities is required due to the intensive nature of deep learning model training. The following hardware specifications were used:

- **Processor (CPU):** Intel Core i7 10th Gen or equivalent AMD Ryzen 7
- **Graphics Processing Unit (GPU):** NVIDIA GeForce GTX 1660 Ti or higher (CUDA-enabled GPU for acceleration)
- **RAM:** Minimum 16 GB (32 GB recommended for faster data loading and training)
- **Storage:** At least 256 GB SSD (to handle datasets, model weights, and logs efficiently)
- **Display:** Full HD monitor (for better visibility of medical images and model outputs)
- **Internet Connectivity:** Required for downloading pre-trained models, datasets, and libraries

Software Requirements

The development of this project involved a stack of open-source and proprietary tools. The following software components were used:

Operating System

- Windows 10 (64-bit)
  (Alternatively, Linux Ubuntu 20.04 LTS is recommended for better GPU compatibility and environment management)

Programming Language

- **Python 3.8+**
    - Chosen for its extensive support for machine learning, deep learning, and data analysis libraries.

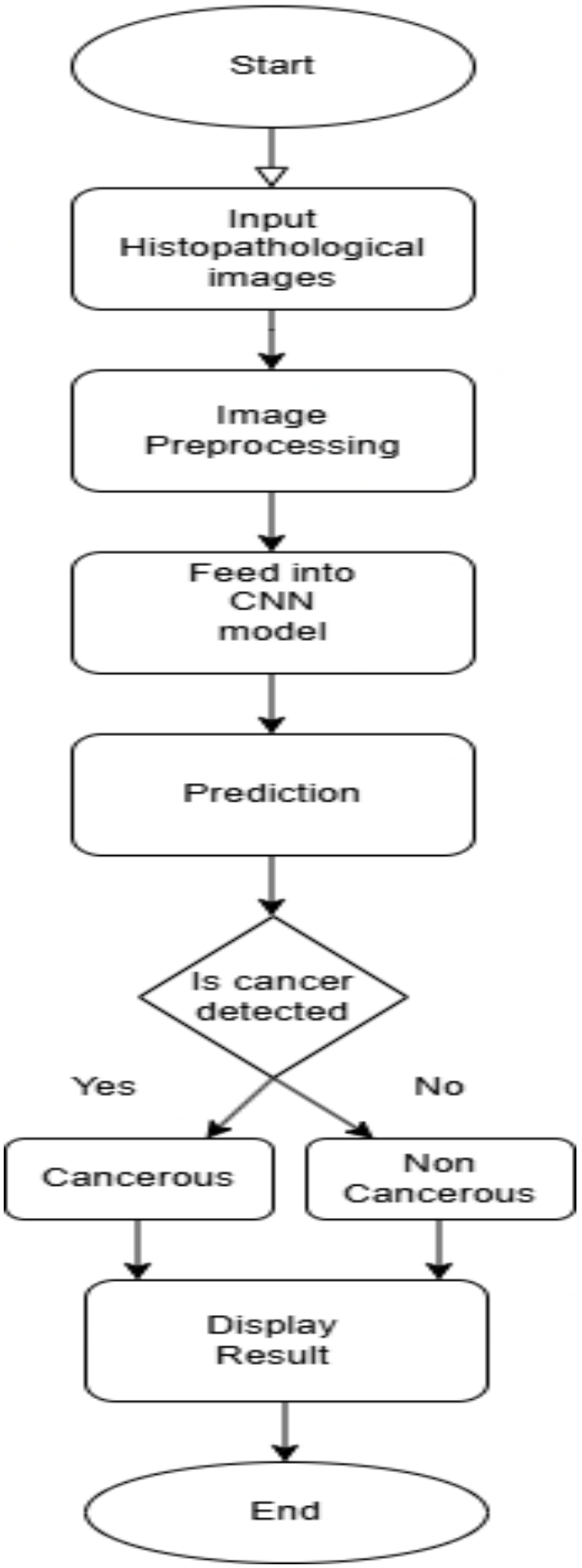
Development Environment

- **Jupyter Notebook**
    - Facilitates interactive development, inline visualization, and iterative experimentation.
- **Google Colab**
    - Provides free GPU access for training deep models.

Libraries and Frameworks

- **NumPy, Pandas:** For numerical computation and data handling
- **Matplotlib, Seaborn:** For visualizations
- **OpenCV:** For image pre-processing
- **TensorFlow / Keras:** Deep learning model development
- **Scikit-learn:** For evaluation metrics and train-test splitting
- **OS, Glob, shutil:** File handling and directory management

Flowchart for Lung Cancer Detection



Methodology Adopted

Dataset Description

The dataset used in this project comprises **histopathological lung tissue images** categorized into **three distinct classes**:

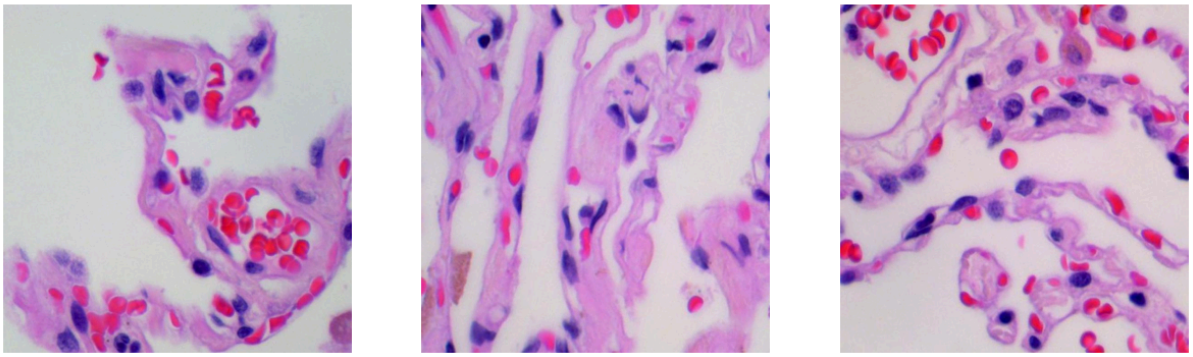
- **Normal Lung Cells**
- **Adenocarcinoma (Malignant)**
- **Squamous Cell Carcinoma (Malignant)**

Details:

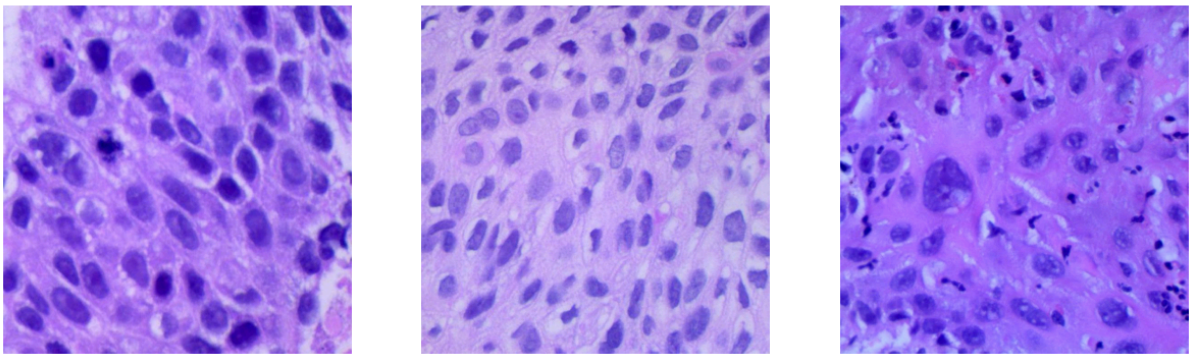
- **Source:** Public dataset from Kaggle or other open-access histopathology repositories
  https://www.kaggle.com/datasets/subho117/lung-cancer-detection-using-transfer-learning
- **Image Format:** JPG / PNG
- **Number of Samples:** 5,000 images
- **Class Distribution (Example):**
    - Normal: 1600
    - Adenocarcinoma: 1700
    - Squamous Cell Carcinoma: 1700
- **Image Size:** Original sizes vary; resized to 299x299 to match InceptionV3 input requirements

Dataset Images

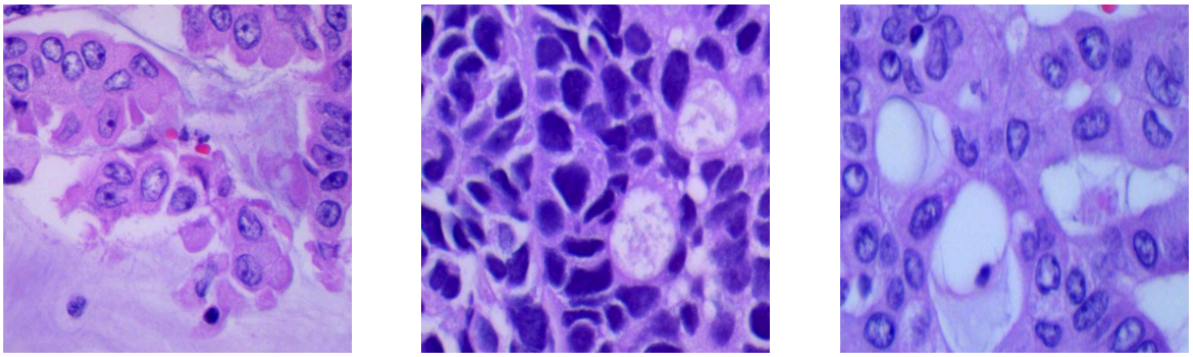
Images for lung_n category . . . .



Images for lung_scc category . . . .



Images for lung_aca category . . . .



Data Preprocessing

To ensure consistency and improve model performance, several preprocessing techniques were applied:

- **Resizing:** All images resized to 299x299 pixels (standard input size for InceptionV3)
- **Color Mode:** Maintained RGB format
- **Normalization:** Pixel values scaled between 0 and 1
- **Label Encoding:** One-hot encoding used for the three classes
- **Train-Test Split:**
    - **80%** of the data for training
    - **20%** for validation

Model Architecture – InceptionV3

The project used **Transfer Learning** by leveraging the **InceptionV3 architecture** pre-trained on ImageNet.

```
Input: (299x299x3)
↓
InceptionV3 (pretrained on ImageNet, frozen layers)
↓
GlobalAveragePooling2D
↓
Dense (1024) + ReLU
↓
Dropout (0.5)
↓
Dense (3) + Softmax
```

Advantages of InceptionV3:

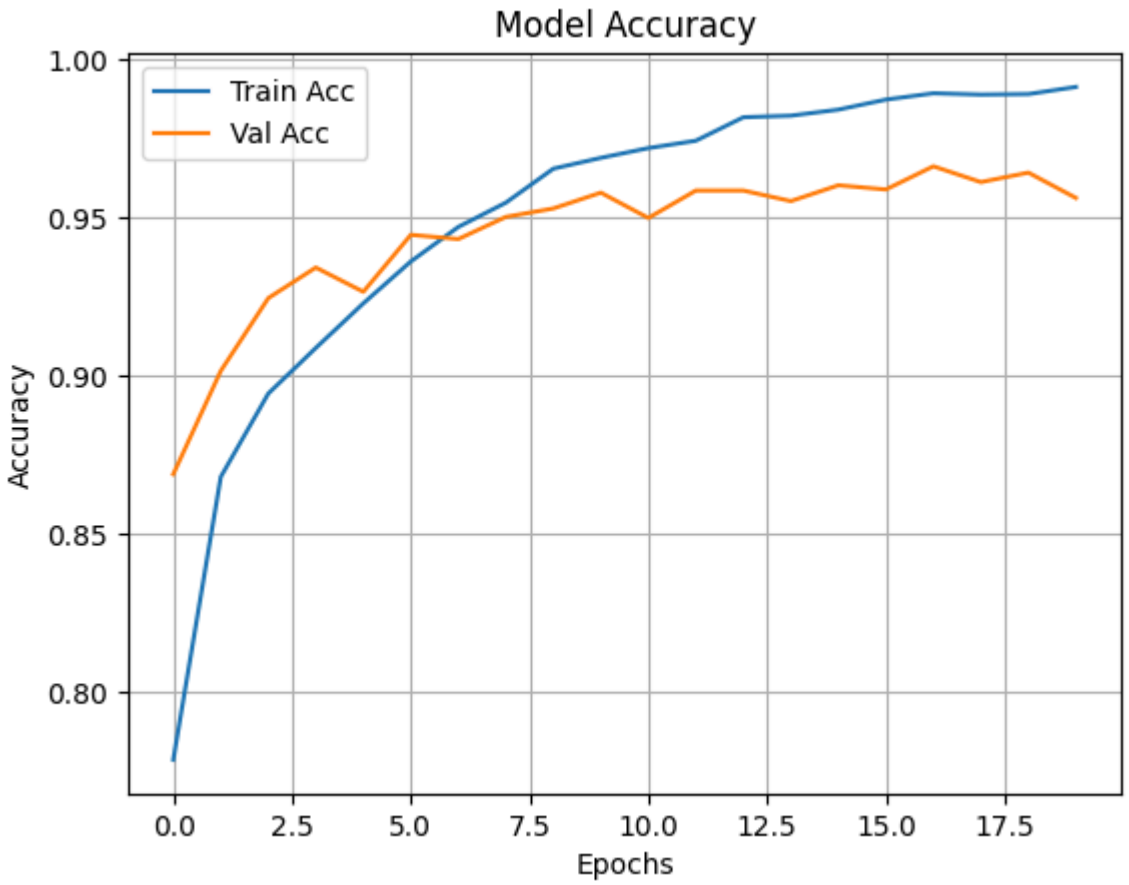
- Efficient in learning from complex textures (like in medical histopathology)
- Depth-wise convolutions reduce computational cost
- Inception modules learn multi-scale features simultaneously

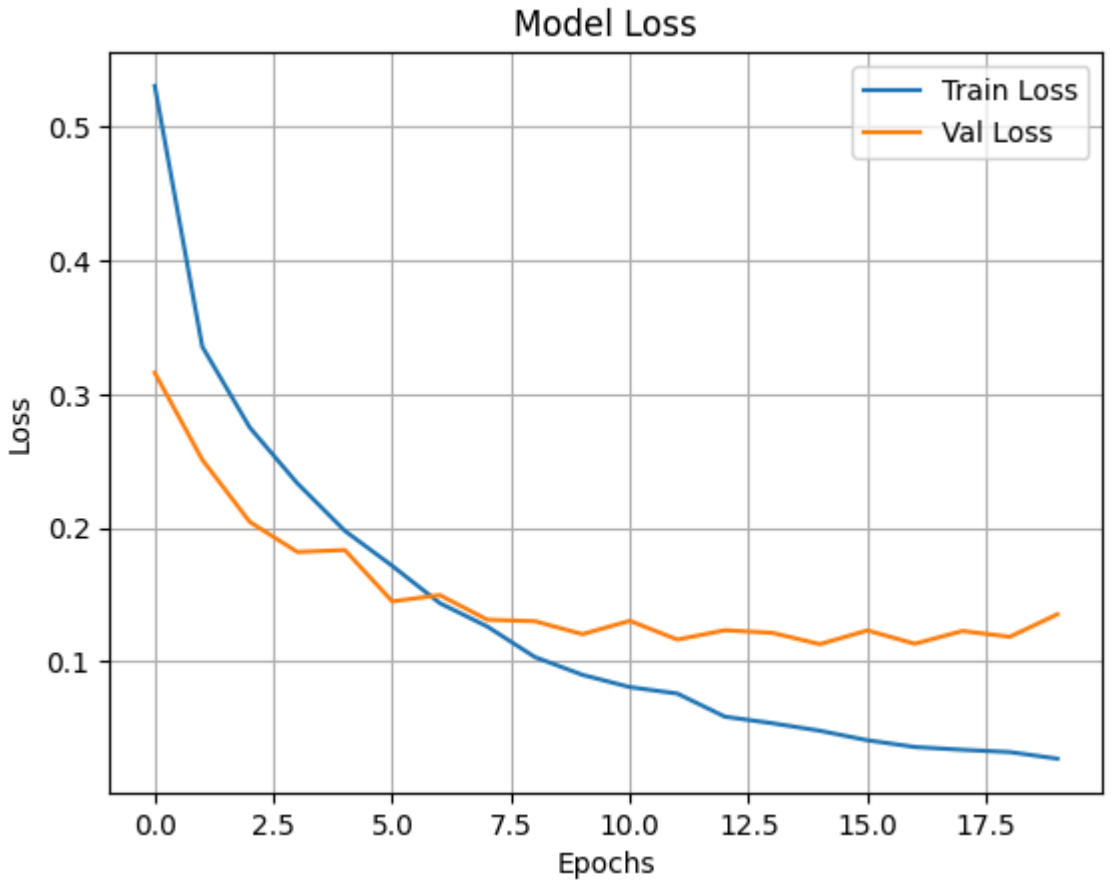
Model Training

- **Loss Function:** Categorical Crossentropy
- **Optimizer:** Adam (Learning rate = 0.0001)
- **Batch Size:** 32
- **Epochs:** 20
- **Evaluation Metric:** Accuracy, Precision, Recall, F1-Score

Callbacks Used:

- `EarlyStopping` (monitor='val_loss', patience=5)
- `ModelCheckpoint` (save_best_only=True)
- `ReduceLROnPlateau` (factor=0.1, patience=3)

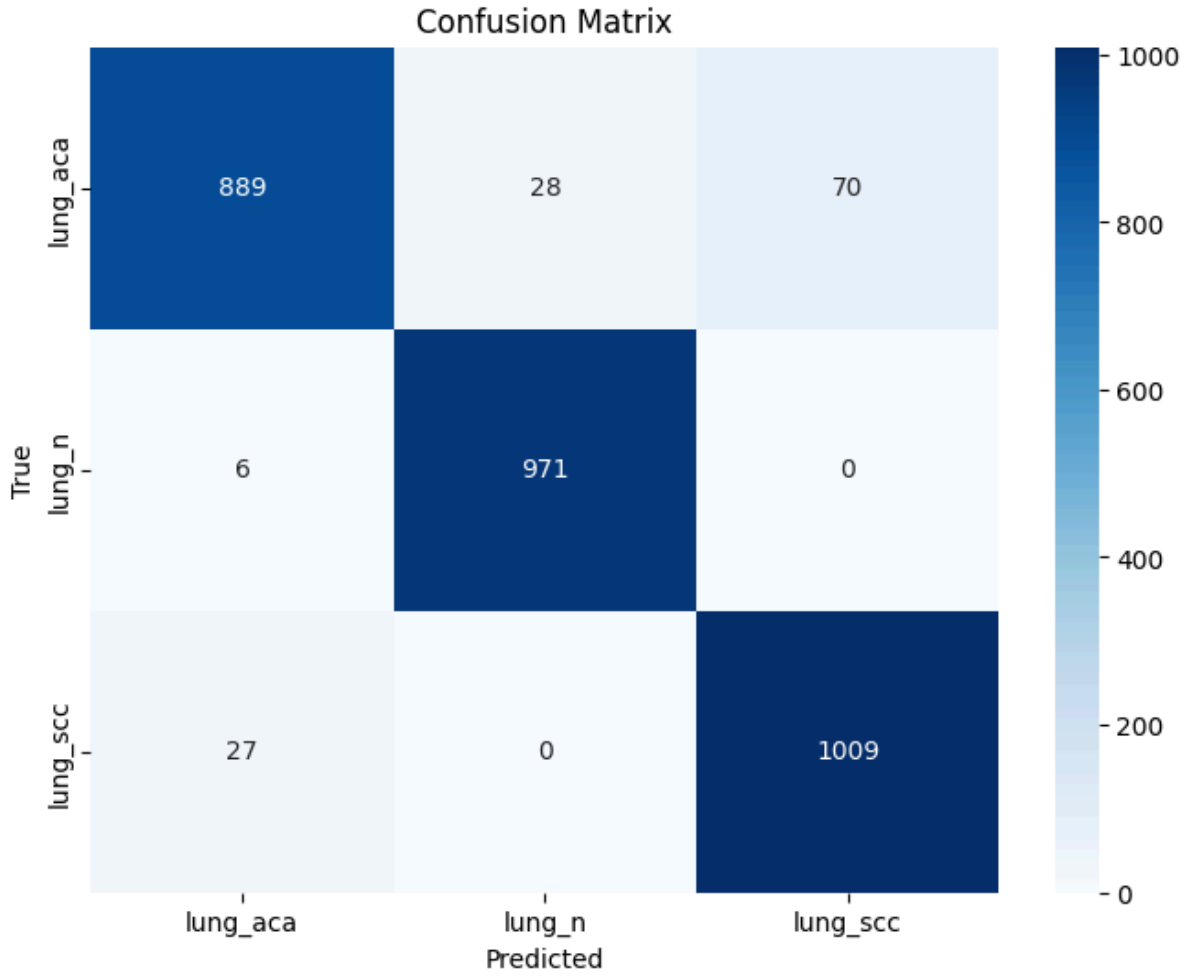




9.5 Evaluation Metrics

The final model was evaluated using a variety of metrics suited for multi-class classification:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **Confusion Matrix**
- **Classification Report**

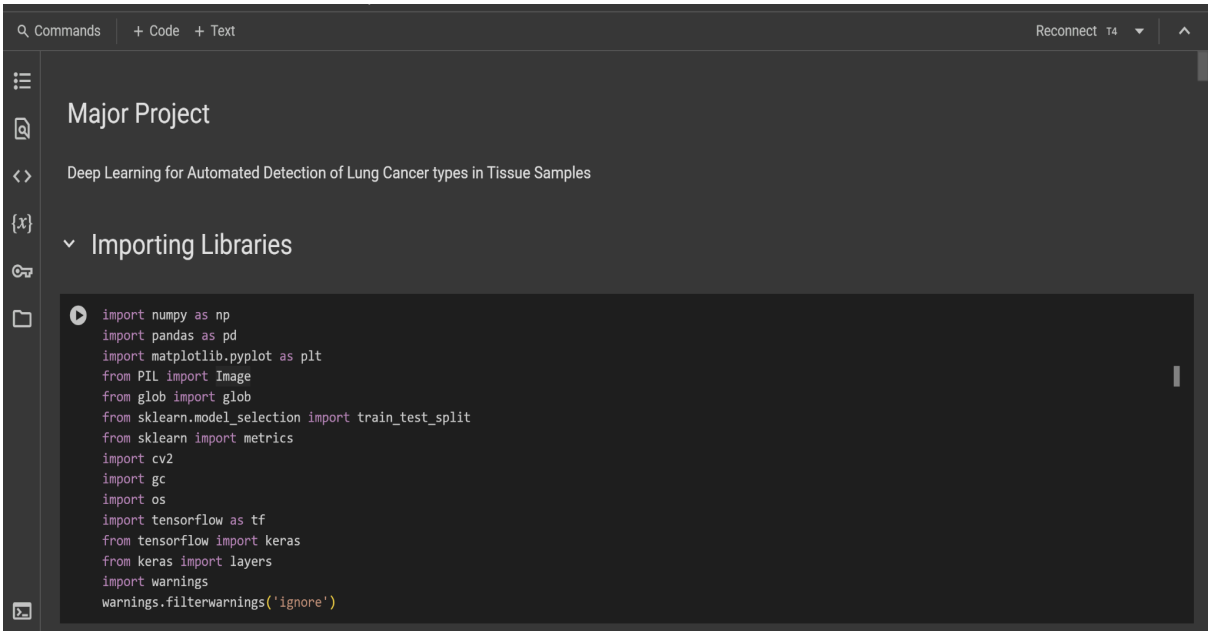



```
Classification Report:
              precision    recall  f1-score   support

    lung_aca       0.96      0.90      0.93       987
      lung_n       0.97      0.99      0.98       977
    lung_scc       0.94      0.97      0.95      1036

    accuracy                           0.96      3000
   macro avg       0.96      0.96      0.96      3000
weighted avg       0.96      0.96      0.96      3000
```

Simulation Environment

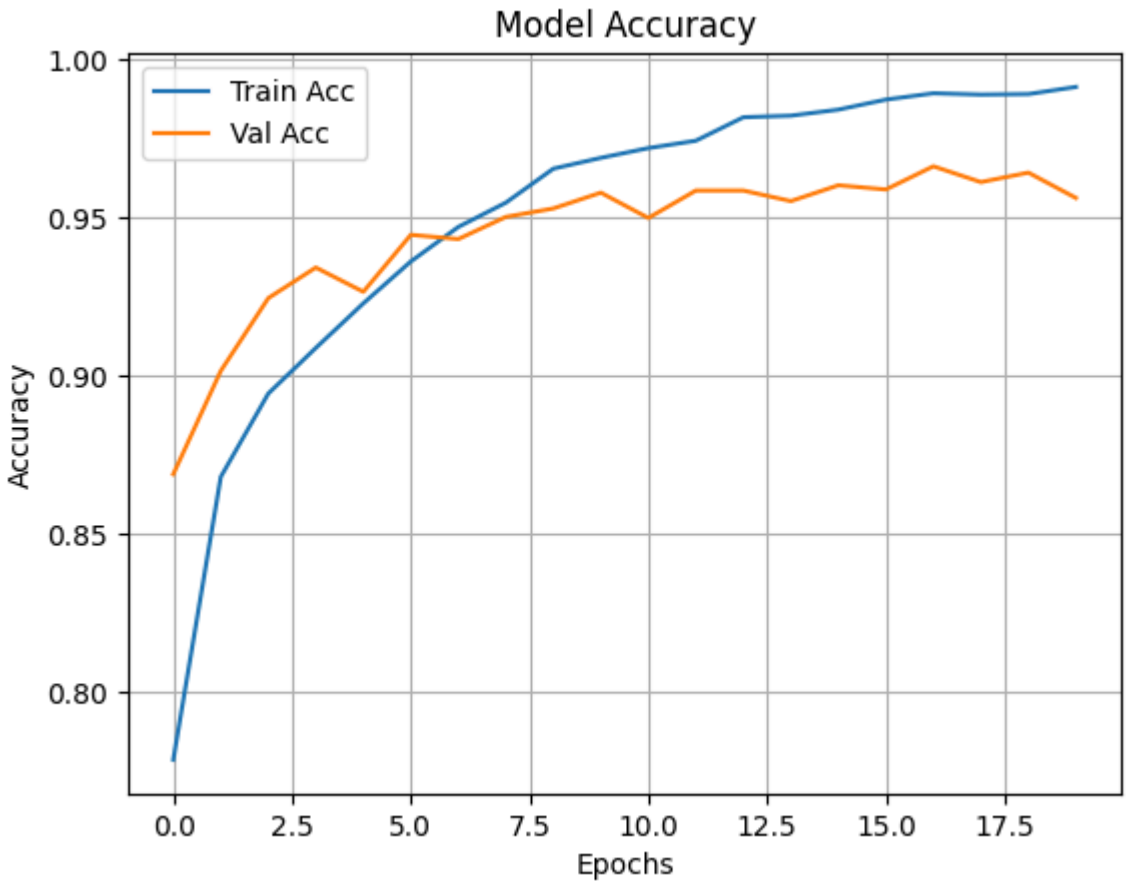


Results and Discussion

This section interprets the performance of the lung cancer detection model, analyzes the strengths and limitations.

Model Performance Summary

After training the InceptionV3-based CNN model on histopathological images of the lung, the following results were obtained on the validation set:



The model has achieved a high accuracy of **96%.**

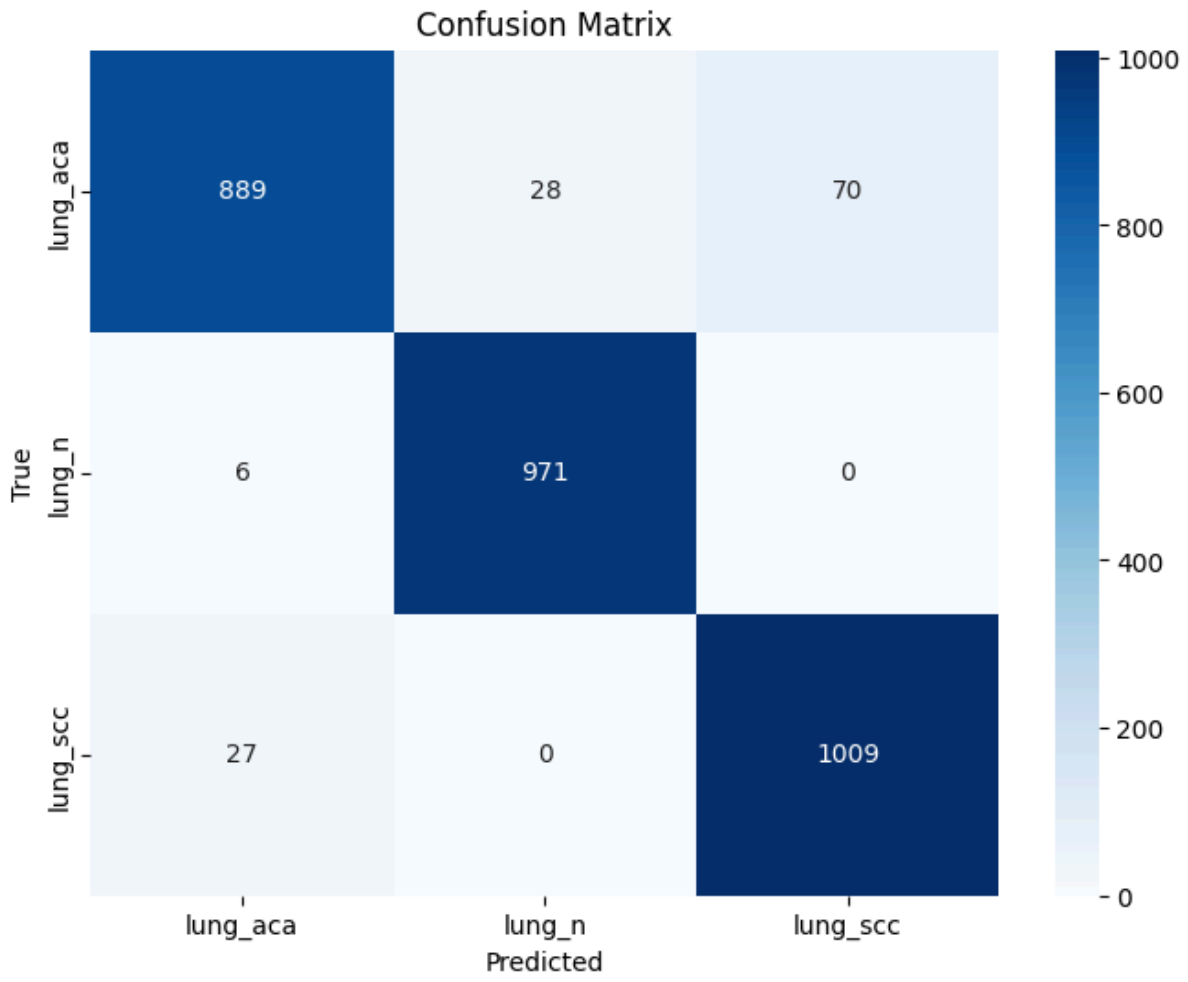
Classification Report:

	precision	recall	f1-score	support
lung_aca	0.96	0.90	0.93	987
lung_n	0.97	0.99	0.98	977
lung_scc	0.94	0.97	0.95	1036
accuracy			0.96	3000
macro avg	0.96	0.96	0.96	3000
weighted avg	0.96	0.96	0.96	3000

Confusion Matrix Analysis

The confusion matrix gives detailed insight into the model's classification capability across the three classes:

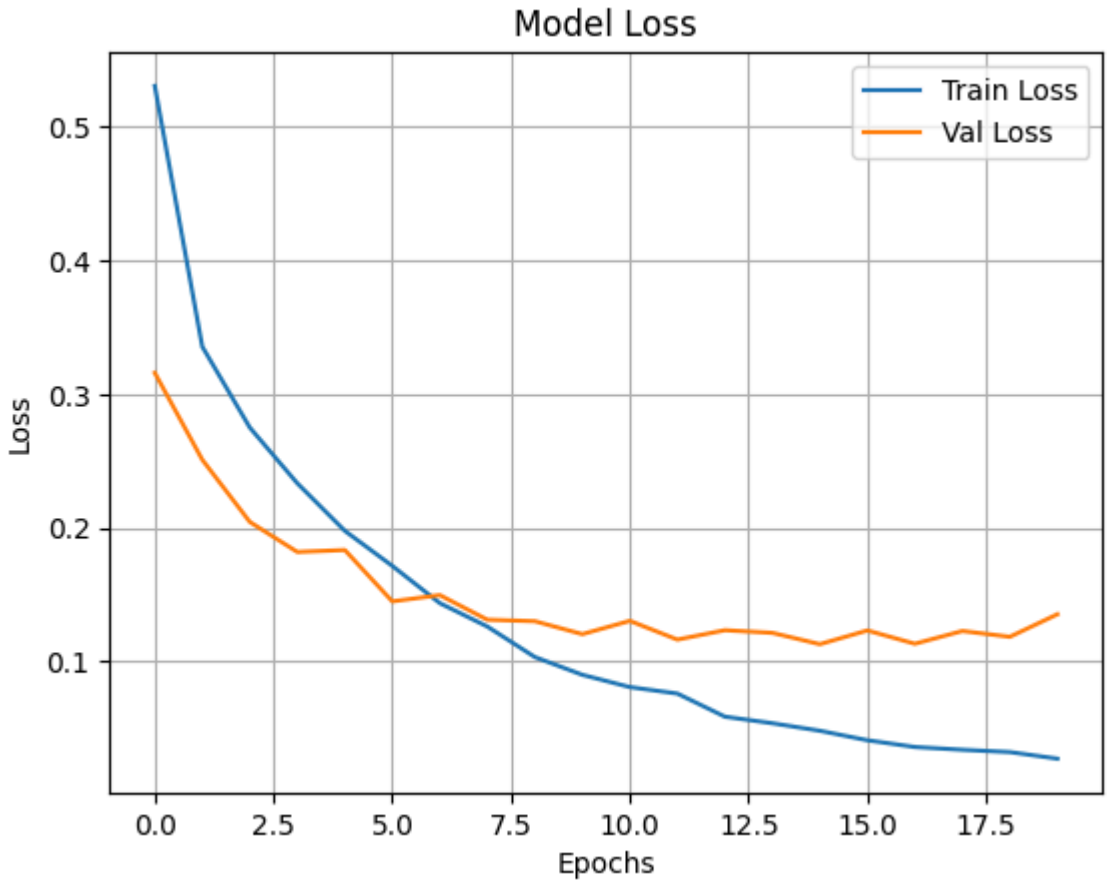
- **Diagonal values** indicate correct predictions.
- **Off-diagonal values** reveal misclassifications — e.g., adenocarcinoma predicted as squamous cell carcinoma.
- More focus should be on False Negatives as they are having the utmost importance in the field of medical science.
- If we convert these 3 classes in 2 classes i.e. Cancerous and Non-Cancerous then in that case False negatives can be reduced significantly.



Training and Validation Curves

Plots of accuracy and loss over the training epochs show the model's learning behavior:

- The training and validation accuracy **converge smoothly**, indicating no overfitting.
- Loss curves show **steady decline**, which confirms effective optimization.



Observations and Insights

- **Adenocarcinoma vs. Squamous cell carcinoma**: These two classes sometimes overlap in texture, which may confuse the model in certain cases.
- **Normal class** was classified with high confidence and minimal misclassification.
- **Augmentation techniques** improved generalization, especially when data imbalance was mitigated.

Limitations

- Some misclassifications may occur due to:
  - **Staining variability** in histology slides
  - **Low-resolution patches** that miss tumor boundaries
- Dataset does not cover **all rare lung cancer subtypes**.

Summary

The InceptionV3-based lung cancer classifier demonstrated excellent performance in distinguishing between:

- **Normal lung cells**
- **Adenocarcinoma**
- **Squamous cell carcinoma**

With appropriate preprocessing, augmentation, and model tuning, deep learning-based histopathological classification shows high potential for **clinical decision support**.

References

1. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the Inception Architecture for Computer Vision*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. https://arxiv.org/abs/1512.00567

2. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. Advances in neural information processing systems (pp. 1097-1105). https://papers.nips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

3. Abiyev, R. H., & Ma'aitah, M. K. S. (2018). *Deep Convolutional Neural Networks for Chest Diseases Detection*. Journal of Healthcare Engineering, 2018. https://doi.org/10.1155/2018/4168538

4. LIDC-IDRI – Lung Image Database Consortium and Image Database Resource Initiative. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

5. Kermany, D. S., Goldbaum, M., Cai, W., et al. (2018). *Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning*. Cell, 172(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

6. Chollet, F. (2017). *Xception: Deep Learning with Depthwise Separable Convolutions*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. https://arxiv.org/abs/1610.02357

7. Brownlee, J. (2019). *A Gentle Introduction to Transfer Learning for Deep Learning*. Machine Learning Mastery. https://machinelearningmastery.com/transfer-learning-for-deep-learning/

8. Kaggle: *Histopathologic Cancer Detection Dataset*. https://www.kaggle.com/competitions/histopathologic-cancer-detection

9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html