PROJECT REPORT

On

Detecting Lung and Colon Cancer Using ResNet on Histopathology Images

Submitted in Partial Fulfilment of Award of

BACHELOR OF TECHNOLOGY

In

Computer Science and Engineering

By

Sanjay S

Prem Kumar S

Alwin Simson P A

Under the Supervision of

Dr. Ezil Sam Leni A

HOD, Department of Computer Science and Engineering

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

ALLIANCE UNIVERSITY

BENGALURU

MAY 2024

Computer Science and Engineering

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

CERTIFICATE

This is to certify that the project work entitled "Detecting Lung and Colon Cancer Using ResNet on Histopathology Images" submitted by Sanjay S [Roll No. 20030141CSE072], Prem Kumar S [Roll No. L200301441CSE107] and Alwin Simson P A [Roll No. 20030141CSE068] in partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Alliance University, is a bonafide work accomplished under our supervision and guidance during the academic year 2023-2024. This thesis report embodies the results of original work and studies conducted by students and the contents do not form the basis for the award of any other degree to the candidate or anybody else.

Dr. Ezil Sam Leni A
Computer Science and Engineering
(Head of Department)

Dr. Ezil Sam Leni A
Computer Science and Engineering
(Head of Department)

External Examiners

1. **Name:** **Signature**

2. **Name:** **Signature**

Computer Science and Engineering

ALLIANCE COLLEGE OF ENGINEERING AND DESIGN

DECLARATION

We hereby declare that the project entitled **"Detecting Lung and Colon Cancer Using ResNet on Histopathology Images "** submitted by me/us in the partial fulfilment of the requirements for the award of the degree of Bachelor of Technology (Computer Science and Engineering) of Alliance University, is a record of my/our work carried under the supervision and guidance of **Dr. Ezil Sam Leni A** , department of Computer Science and Engineering

We confirm that this report truly represents the work undertaken as a part of our project work. This work is not a replication of work done previously by any other person. We also confirm that the contents of the report and the views contained therein have been discussed and deliberated with the faculty guide.

Name of the Student	University Registration Number	Signature
Prem Kumar S	**L20030141CSE107**	
Sanjay S	**20030141CSE072**	
Alwin Simson P A	**20030141CSE068**	

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We are very thankful to our project guide **, Dr. Ezil Sam Leni A** , Head of the department and Professor of department of Computer Science and Engineering, for her sustained and inspiring guidance, along with her cooperation throughout the project. Her wise counsel and valuable suggestions have been truly invaluable.

We would like to thank **Dr. Reeba Korah**, Dean - ASAE, for their encouragement and cooperation at various levels of Project.

We avail this opportunity to express my deep sense of gratitude and hearty thanks to the Management of Alliance University, for providing world class infrastructure, congenial atmosphere, and encouragement.

We express my deep sense of gratitude and thanks to the teaching and non-teaching staff at our department who stood with me during the project and helped me to make it a successful venture.

Prem Kumar S

Sanjay S

Alwin Simson P A

Abstract

Detection of lung cancer and colon cancer in histopathology images using ResNet is very important for early and accurate diagnosis, which is a prerequisite for effective treatment. Recent advances in deep learning, especially CNNs, have demonstrated excellent performance in analysing medical images, thus providing a better strategy for detecting blood-eating tumours. Preprocessing of histopathological images to improve the ability to identify malignant areas is dependent on the plan.

Then video extraction and image classification are based on ResNet architecture. Histological images of lung and spinal cord tissue annotated with benign and malignant data will be used to train the ResNet model. This will help the model understand what the tissues and cells look like so it can determine where the cancer is in the lung and colon. A comprehensive evaluation will lead to independent data that will verify the accuracy, sensitivity, and specificity of the plan. Comparison with existing methods will demonstrate the robustness and efficiency of the ResNet-based method for diagnosing lung disease and cancer from histopathology images. This study shows how deep learning, specifically ResNet, can help improve doctors' ability to detect lung and breast cancer at an early stage. With timely intervention and treatment planning, this recommendation may improve patient outcomes. But more research and clinical trials are needed to translate this into clinical practice and help develop more effective drugs for cancer and cancer patients. Ongoing research is needed to improve the ResNet-based approach and ensure its effectiveness in real-world situations in lung and cancer treatment.

List of Figures

LIST OF TABLES

LIST OF ABBREVIATIONS

CT scan	Computed Tomography Scan
LDCT	Low dose Computed Tomography
CNN	Convolution Neural Network
ResNet	Residual Network
ReLU	Rectified Linear unit
CAD	Computer-aided Design
CNN-ALCD	Convolution Neural Network Automatic Lung Cancer Detection
CDSS	Clinical Decision System
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
SGD	Stochastic Gradient Descent
$Accu_y$	Accuracy
$Prec_n$	Precision
$Reca_l$	Recall

List of Equations

Sl. No	Equations	Page No
1	Precision	31
2	Recall	32
3	Accuracy	32
4	F-Score	32

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Although ResNet-50 has excellent performance in diagnosing lung and lung cancer from histopathology images, this approach has significant limitations. First, these deep learning models rely heavily on registry data, which is often difficult to obtain, especially for rare or special types of cancer.

Additionally, results for deep learning models such as ResNet-50 are difficult to interpret because their inner workings may not be easily understood by medical professionals. This will reduce the reliability and validity of the procedures used in the clinic. ResNet-50 and other similar models may have issues affecting the external model of distribution of training data; this can lead to it being misclassified or missed, especially in the case of an unusual or rare cancer.

Examining patterns in a picture is fine, but it doesn't capture social relationships or contextual information important for accurate diagnosis of cancer. Histopathological analysis often includes information about the distribution of cancer cells in tissue samples, and this can be better accomplished with advanced techniques than traditional neural methods.

1.1 Introduction to Lung and Colon Cancer Detection

Detection of lung and colon cancer is important for early intervention and better prognosis. Both lung and colon cancer are at the top of the list with the most cancer deaths. Detection of breast cancer is done by screening, though it usually is asymptomatic. The article has elucidated some of the methods used in the diagnosis of lung and lung cancers. It calls for the seriousness of screening, detecting, and molecular testing in the prognosis. The risk factors include current or past smokers and individuals exposed to asbestos at their workplaces. Low-dose computed tomography screening has been effective in the reduction of the cancer incidence rate among the high-risk population. It is done by exposing the lung to multiple cross-sectional images that require minimal radiation, hence spotting small or large nodules that may indicate the presence of cancer. Be aware of the signs and symptoms and get some diagnostic tests done, including chest X-ray, computed tomography scan, and positron emission tomography scanning. Chest X-rays often are used as initial screening tools, although their sensitivity in detecting early-

stage lung cancer is low. CT scans demonstrate a high sensitivity and specificity, providing good anatomic information and allowing for a characterization of the lung. PET scans reveal areas of metabolic activity, help distinguish benign from malignant tumours, and measure disease spread.

During the past few years, molecular and genetic testing for the diagnosis and treatment of cancer, particularly small cell lung cancer, has become an integral component. Biomarker analysis of the tumour—for example, mutations of the epidermal growth factor receptor gene, anaplastic lymphoma kinase gene rearrangements, and expression of PD-L1—identifies those patients who would benefit from medical or immunological indications. Molecular targeted therapies are revolutionizing the treatment of lung cancer and tailor treatment to individual patients, thus enhancing patient outcomes and quality of life.

1.2 Introduction to ResNet Model

ResNet is a very powerful neural network, mostly used as a convolutional neural network, which helps in diagnosing lung cancer from histopathology images. Such depth would allow the network to learn through the hierarchical structures of input images, an indispensable prerequisite for capturing complex patterns found in histopathological images showing lung cancer. The main innovation of ResNet is the use of redundant blocks. Each of the residual segments has a cross-connection, otherwise known as a fast connection, that enables the network to pass one or more layers. This solves the vanishing problem that could happen with a relationship that is too deep. For instance, in the case of diagnosing lung cancer by histopathology images, ResNet can be trained for the structure and properties of cancer tissues, which may include many irregular shapes in cells, abnormal changes of cells, or specific patterns associated with growth in cancer.. Large file of Rigaku images. The model is trained via supervised learning; it learns to classify images into relevant labels, like cancerous or non-cancerous images, by adjusting internal parameters during learning. Once the ResNet model is trained, it can be run on individual images to infer its performance in diagnosing lung cancer. Model performance metrics can be used to judge its performance, such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve.

Therefore, the fact that greater accuracy has been achieved with respect to previous research data marks a great achievement and shows that our way of diagnosing lung cancer using histopathology images works. This better accuracy is a statement of the use of the ResNet network architecture, which is state-of-the-art in the convolutional neural network architecture,

for detecting the subtle patterns of lung cancer in histopathology samples. Our research has yielded good results not only in confirming the effectiveness of machine learning in diagnosing lung cancer but also in a positive way to support pain treatment. This may have significant implications for patient care and outcomes in the field of lung cancer screening by expediting the diagnostic process and reducing the possibility of misdiagnosis.

ResNet50 is a network with only 50 layers. It is a deep neural network architecture mainly used in computer vision tasks, especially for image classification, target detection and image recognition. It was developed by Kaiming He et al. in 2015. It solves the challenge of training deep neural networks.

The main innovation behind ResNet50 is that it uses residual learning or shortcut connections, which allows for deep network training without running into the problem of loss. With the number of layers being increased in normal neural networks, the slope of the loss function will get decreased during the recovery period. ResNet50 overcame this limitation with the help of cross-learning by learning the remaining map, which can be interpreted as the difference between the input and the output of a given layer.

The architecture of ResNet50 involves a couple of building blocks: convolutional layer, batch normalization, activation, maximum pooling layer, and residual layer. This network has 48 convolutional layers in 16 residual blocks with additional layers for pre- and post-processing. The first layer of ResNet50 exploits techniques such as convolution, batch normalization, and rectified linear unit to extract features from the input image. This layer is responsible for training deep links. Each part of the architecture has a set of convolutional processes followed by batch normalization and ReLU activations with fast connections spanning one or more layers. These fast connections enable the network to learn the rest of the graph, hence making the gradient flow easier during training and alleviating the vanishing gradient problem.

ResNet50 exploits a short self-connection to implement cross-connections, which add input into the output of the residual block and, thus, effectively bypass the convolution process. This enables the network to learn the rest of the map instead of trying to learn what to do directly on the map; hence, the job of deep networks becomes easier and perfects tasks such as image classification. The final process normally includes global mean pooling, a fully connected layer, and SoftMax activation, which transforms the captured features into probability

categories for the image distribution function. In this way, the network outputs the most likely classes, given the input image in the final SoftMax layer during inference.

CHAPTER 2

Literature Survey

2.1 Literature Review

R. Pandian, V. Vedanarayanan, D.N.S. Ravi Kumar, R. Rajakumar, "Detection and Classification of Lung Cancer Using CNN and Google Network", ScienceDirect, Page 2022.

An algorithm using deep learning tools applied to lung cancer diagnosis. . The core network chooses VGG-16 architecture. This study used deep learning to distinguish lung cancer (e.g., adenocarcinoma, cellular carcinoma, squamous cell carcinoma) from lung images. Apply two different pre-trained neural networks (such as GoogleNet and Vgg16 network) to the data to determine their performance (such as accuracy, precision, and significance) and compare these results with the CNN network.

Comparing the two layers deconvolution layers are compared with a normalization layer and a pooling layer. 100 sample images are taken from each group to train and evaluate the performance of the network. 70 of these images are used for training and the remaining 30 for validation. It turns out that GoogleNet and Vgg16 networks achieve accuracy using only two CNNs.

Therefore, GoogleNet and Vgg16 networks are the first choice. GoogleNet has a depth of 28 layers, a memory size of 27 MB, 7 million parameters and accepts images with a size of 224x224 pixels. Vgg16, on the other hand, has 16 layers, a small memory of 535 MB, and accepts image sizes of 224x224 pixels. CNN algorithm and Googlenet were chosen to detect cancer areas and classify them as normal and abnormal. Deep convolutional network architecture called VGG16 is used as a simple network to implement the CNN algorithm. The proposed algorithm effectively identifies lung cancer.

Hamed Alqahtani, Eatal Alabdulkreem, Faiz Abdullah Alotaibi, Mrim M. Alnfiai, Chinu Singla and Ahmed S. Salama, -Improved water srider algorithm using convolutional autoencoders for diagnosis of lung and gastrointestinal diseases in histopathology images — 12,EE, volume.

This IWSACAE-LCCD system is designed to identify and classify cancer cells as well as cancer cells. The IWSACAE-LCCD method includes MF-based preprocessing, MobileNetv2

feature extractor, IWSA-based hyperparameter tuning, and CAE-based cancer diagnosis. Application of the IWSA model helps guide the selection of hyperparameters relevant to the CAE algorithm that can be used to identify lung and colon cancer occurrence. The simulation method was used to improve the detection of IWSACAE-LCCD technology. The results show that the IWSACAE-LCCD system achieves better results than other systems. Future work will focus on accessing higher quality HI data, which is necessary for training and implementing deep learning models. Collecting data in cancer and non-cancer settings can be challenging, and inconsistent data can lead to poor performance standards. Moreover, deep learning models have high demands in terms of budget to achieve high HI solutions. Or it can be used with existing systems.

Nusraat Nawreen, Umma Hany, Tahmina Islam, "Diagnosis and classification of lung cancer using CT scan images", IEEE, p. 2021

Computed tomography (CT) is the most popular and best way to diagnose lung cancer. However, visual interpretation of CT scan images is difficult, time-consuming, and can lead to misinterpretation of malignancy. Therefore, computer aided technology is needed to accurately diagnose lung diseases. There are many methods available in the literature.

In this paper, we propose a new method to detect and classify cancer cells by CT scan imaging. We use different presets to soften and enhance the image. We then used index and edge measurements to segment the lung tumour region of interest (ROI). Finally, we calculated the geometry of the extracted ROIs and classified them into weighted and weighted weights using a support vector machine (SVM) classifier. We found significant accuracy in diagnosing lung cancer nodules and estimating body weight using our proposed method.

Binson; M Subramaniam; G.K. Ragesh; Ajay Kumar, "Early Detection of Lung Cancer Through Breath Analysis using AdaBoost Ensemble Learning Method", IEEE PP.2022.

An electronic nose machine based on an iron oxide generator was developed to identify and modify organic compound biomarkers in exhaled breath. The system was tested on 10 cancer patients and 15 healthy controls to distinguish their breathing patterns. Using independent screening tools, the system achieved accuracy, sensitivity, and specificity of 76%, 70%, and 80%, respectively. More research is needed to better understand the function of these systems in the early detection of cancer, as in the early stages of cancer.

Kanakala Raja Sekhar, GRL M Tayaru, Antharaju K Chakravarthy, Boyina Gopiraju, A. Lakshmanarao, "Pain Management Study Using Convolutional Convolution and Neural Networks", IEEE, p. 2024.

Cancer is the world's leading health problem and death. Lung cancer is one of the leading causes of cancer worldwide, and early detection requires new methods to improve diagnosis and subsequent treatment. Machine learning, as a form of artificial intelligence, provides a clear way to improve the reliability and efficiency of lung cancer diagnosis.
The aim of the project is to use the power of deep learning to create an early and effective cancer diagnosis tool by analyzing CT scan images. We conducted an experiment using Kaggle's cancer database. We use CNN and Resnet50 technology to make appropriate cancer diagnosis. The results showed that the proposed model had good accuracy compared with the lung cancer diagnosis model.

Mattakoyya Aharonu, R Lokesh Kumar, "Convolutional neural network framework for automatic detection of lung cancer based on lung CT images," IEEE, p. 2022.

Lung cancer is one of the leading causes of death worldwide. In today's medical development, chemotherapy is widely used to treat lung cancer. However, research on early diagnosis of lung cancer is important in saving people's lives. With innovations in machine learning, computer-aided design (CAD) systems that can diagnose cancer have entered mainstream medicine. In particular, deep learning models such as convolutional neural networks (CNN) have proven to be effective methods for learning features from computed tomography (CT) scan images and tests showing cancer outcomes. In this paper, we propose a CNN-based model for lung cancer diagnosis from lung CT scan images. We propose an algorithm called CNN-based automatic lung cancer diagnosis (CNN-ALCD), which is based on learning analysis. This working model enables the diagnosis of lung cancer with a new diagnostic model. Strategies to solve the problem include various methods such as preprocessing, creating different CNNs, training CNN models, and performing cancer detection. Experimental results show that the proposed CNN-based model successfully implements multiple neural network methods with an accuracy of up to 94.11%. Thus, the proposed system can be integrated with the hospital's clinical decision system (CDSS) to provide automatic lung disease diagnosis.

Shireen Al-Ofary, Hamza Osman İlhan, "PCA-based simple deep features classification of SVM for lung and cancer diagnosis", IEEE, p. 2023.

In this case, the vehicle creates intelligence-based tools during the diagnosis phase. In this study, our deep learning method was used to detect cancer images and lung cancer in our case. AlexNet, SqueezeNet and ShuflleNet are used as default models in the transformation. In the first scenario, CNN was used as a classifier for cancer and cancer images. The classification performance of each model is reported separately as SoftMax performance. In the second step, only the features obtained from the images obtained from the pre-training samples are collected and fed to the SVM classifier. In the last scenario, PCA is applied to features extracted from the network to reduce the number of features. Diagnosis of cancer and lung diseases by third-party methods achieved the highest results. Specifically, using ShuffleNet, the classification of breast cancer and lung cancer is 99.93% and 97.92%, respectively.

2.2 Limitations of the Existing System

i. **False positives and negatives:** Low-dose CT scans, a common screening tool, can mistakenly suggest cancer (false positive) or miss it entirely (false negative). This can lead to unnecessary worry or a delay in treatment.

ii. **Radiation exposure:** CT scans, often used to screen and diagnose lung cancer, expose patients to ionizing radiation. Although the radiation dose of a CT scan is lower than a standard CT scan, there are still risks associated with increased radiation exposure, especially for people who have a repeat check. This is especially true for high-risk groups such as current and former smokers, who may be screened regularly as part of blood-consuming cancer screening.

iii. **Overdiagnosis:** Lung cancer screening tests can reveal slow-growing or slow-growing tumors that may not be dangerous if left untreated. However, detection of these tumors can lead to overdiagnosis, where people are diagnosed with cancer that does not cause symptoms or reduces their prognosis. This can lead to overtreatment, including unnecessary surgery, chemotherapy, or radiation therapy, putting patients at risk without receiving medical benefits. Overdiagnosis and overtreatment not only burden patients but also impact medical resources and increase medical costs.

iv. **Limited applicability:** One of the main limitations of cancer diagnosis is the limited use of certain research methods or techniques. Although technologies such as CT scans are widely used for cancer screening and lung cancer screening, they may not be suitable for everyone, especially those with certain medical conditions or body limitations. For example, a person

with severe claustrophobia may have difficulty undergoing a CT scan, and a person with kidney dysfunction may be at risk for contrast nephropathy. Similarly, some diagnostic procedures, such as bronchoscopy or needle biopsy, may not be possible for people with respiratory disease or anatomical limitations. This limited use will lead to inequalities in access to timely and appropriate cancer screening, especially for vulnerable groups.

2.3 Limitations in Current System

i. **False positives and negatives:** A low-dose CT scan, since it is a screening tool applied as a matter of routine, would probably tend to over-detect cancer or miss it altogether. This might cause unnecessary worry or delay in treatment.

ii. **Radiation exposure:** CT scans often are used for both screening and diagnosis of lung cancer; they involve exposure to ionizing radiation. Though the amount of radiation in a CT scan is usually lower than a diagnostic CT scan, there is, of course, some risk of increased radiation exposure, especially to people who undergo repeated tests. Sufferers of this are high-risk groups, including current and former smokers, who are screened regularly, part of blood-guzzling cancer screening.

iii. **Overdiagnosis:** Tests for lung cancer screening can pick up slow-growing or growing tumours. These are tumours that, if left alone, would not cause a problem. Still, the detection of these tumours can cause overdiagnosis. It means that people are diagnosed with cancers that will not cause symptoms or shorten their prognosis. This may result in overtreatment: needless surgery, chemotherapy, or radiation therapy poses risks to the patient without benefiting the patient in medical terms. Overdiagnosis and overtreatment affect not only the patients but also the medical resources and increase the costs of medicine.

iv. **Limited applicability:** Probably the most significant constraint to cancer diagnosis is the limited applicability of certain research methodologies or techniques. While technologies such as CT are already extremely popular in the general sense for cancer screening and screening of lung cancer, most of them are not applicable in all sorts of people, specifically not for those suffering from certain conditions within the body. For example, a person with severe claustrophobia will not be able to go for a CT, while a person with impaired kidney function exposes him to the risk of contrast nephropathy. Similarly, some diagnosis procedures, such as bronchoscopy or needle biopsies, cannot be performed on patients suffering from respiratory disease or anatomic constraints. This limited applicability will lead

to unequal, timely, and appropriate access to cancer screening, particularly for vulnerable groups.

v. **Limited Sources of Data:** The next challenge related to the diagnosis of lung cancer is the limited and heterogeneous pool of data for analysis and decision making. Even though these imaging studies are important for providing information on the presence and characteristics of pulmonary embolism or masses, they may not describe other clinical or molecular characteristics that would lead to better diagnosis and prognosis. Moreover, integrating information from a number of different sources, such as electronic medical records, image archives, and molecular repositories, is problematic as information entry, storage methods, and privacy policies vary.. Limited availability of comprehensive and integrated data may hamper the creation of robust predictive models and decision support tools for diagnosis and control of lung cancer.

vi. **Recency Bias:** Recency bias is the tendency of giving priority or prominence to the most recent information or experience and neglecting past information. Systemic bias in diagnosing lung cancer could even affect a doctor's view of blood tests or biopsy, leading to wrong diagnosis or mistakes. For instance, a radiologist could focus on new lung nodules on a CT scan while totally overlooking the other nodules identified previously, which could be stable but may be just as important. Likewise, physicians may be struck by the new characteristics of the newly obtained biopsy samples and overlook the significance of clinical or radiological findings, which may compromise the accuracy of the diagnosis of the medical examination. Persistent bias will impact the reliability and validity of the methods of lung cancer screening and underline the significance of physical and objective examination that considers all clinical data and pattern.

vii. **Availability of resources:** This is another limitation found in research into lung cancer. The researchers' access to the databases and other resources can be hampered by factors such as school documents, registration fees, and addresses. To perform data analysis and synthesize evidence already existing for clinical studies and opinions, access to databases like PubMed, Scopus, or Web of Science is necessary. However, for resource-poor settings or for researchers who are affiliated with institutions that do not have good libraries, the limitation of access to such archives may make it difficult for one to trace relevant studies and remove relevant information.

viii. **Quality assessment:** Problems that have cropped up regarding the assessment of study quality are variability in study design, sample size, and methodologies. Limited Clinical Validation.

ix. **Emerging research:** Part of the challenges in diagnosing lung cancer is that there is rapid progress and new trends in science and technology. The field of lung cancer diagnostics continues to evolve with the development and application of new imaging modalities, biomarkers, and technologies. But as part of the diagnostic process, data analysis does not capture new advances or new trends. This is because commercial advertising, delayed release of research, or the limitation of peer-reviewed research to certain research areas.

2.4 Scope of The Project

Data Collection and Preprocessing: A wide dataset shall be collected regarding medical images; it might be the CT scan of lungs or the X-ray image regarding the diagnosis for the presence or absence of lung cancer. Basic preprocessing shall be done, including the normalization of images, resizing of images, and augmentation for the sake of enhancing the performance and generalization of the model.

Model Selection and Architecture: For the selected problem, a suitable neural network architecture can be selected from ANN and ResNet because of efficiency in handling image data and extracting the relevant features. It will try different configurations of layers with different depths and hyperparameters for optimizing the performance of the model.

Training and Validation: Use the dataset to train the selected models and apply leading methods such as cross-validation in an attempt to estimate model generalization and avoid overfitting. Metrics such as loss and accuracy can be monitored during training to get a general feel for model convergence and performance.

Evaluate and Fine-Tune: Estimate the performance of the trained models on independent test datasets for their accuracy of detection in lung cancer. Model parameters and architecture are then fine-tuned based on the results coming from the validation data to further improve the accuracy and robustness of the model.

CHAPTER 3
System Design

3.1 Problem Definition

ResNet in the detection of lung and colon cancer through histopathology imaging makes use of deep learning models, majorly ResNet50, in classifying histopathology images of lung and colon tissue as cancerous and non-cancerous cells. With histopathology data derived from lung and colon tissue biopsies, the goal shall be developing ResNet50-ConvNets capable of identifying and classifying regions that bear features indicative of cancerous pathology.

This model is to distinguish between benign tissue and the various stages of cancer, such as malignant tumours and metastatic lesions, that affect lung and colon samples. By evaluating the microscopic attributes present in the histopathology images, the ResNet50 model would identify patterns associated with the advancement and metastasis of the cancer, thus enabling it to classify cancerous and non-cancerous areas accurately.

A well-trained ResNet50 model generalizes from annotated histopathology datasets to cancer pathology across heterogeneous tissue samples. This trained model then becomes a very effective tool in assisting pathologists and health professionals in making quick and exact diagnoses related to lung and colon cancers from histopathology images.

The use of ResNet50 and deep-learning technology will be harnessed in this project to enhance precision, effectiveness, and consistency in diagnosing cancer for better patient outcomes and timely interventions for patients affected by cancer in lungs and colons.

Key Components of the problem:

Data Collection and Preprocessing: Collect an enormous database of images in histopathology, considering representative lung and soft tissue biopsies, including cancer and non-cancerous samples. The data should therefore capture all tissue types, sizes, and histological features for the proper training of models and generalizations. Preprocess the histopathological images to enhance image quality by removing noise and standardizing image sizes and densities. Some of the processes may be normalization, resizing, cropping, and other techniques to improve model performance and flexibility against changes in image properties. The ResNet50 model architecture will be used for this deep learning model in the extraction and classification of features. It demonstrates great performance for image recognition and

good network connectivity that helps benefit the training of deep neural networks. This architecture also alleviates the light gradient problem.

The ResNet50 model architecture will be used for this deep learning model in the extraction and classification of features. It demonstrates great performance for image recognition and good network connectivity that helps benefit the training of deep neural networks. This architecture also alleviates the light gradient problem.

As such, the evaluation of ResNet50 has to be based on accuracy, precision, recall, F1 score, and area under the receiver operating characteristic. Test the model for sensitivity with respect to a few types and stages of cancers, and its ability to identify cancerous and non-cancerous samples.

3.2 System Architecture

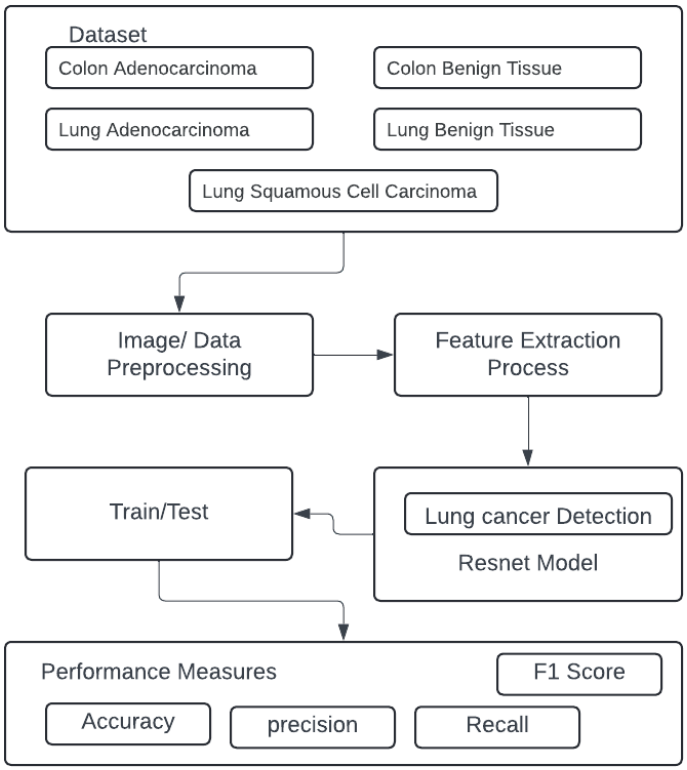

Figure 3.1. System Design of Lung and Colon Cancer.

The system initiates with the imaging dataset of lung and colon histopathology. These images can have numerous kinds of lung and colon cancer, together with cancerous cells and softtissue,

which have their origin in lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and soft tissue benign.

The main image pre-processing deals with resizing the image, converting it to a specific format, or normalizing the density of the pixels. Feature extraction reduces the image to those features that are most important in the classification process. For a diagnosis of lung cancer, these might relate to the size, shape, and texture of the lung nodule or mass. The training method is used to train the ResNet model, whereas the testing method is used for the model performance evaluation. This is also a common problem in deep neural networks where the model is trained by itself to learn what features represent cancer. The model inputs new images to derive a classification of whether those images contain cancer or not.

3.2.1 System Architecture and Explanation

3.2.1 Dataset

Histopathology images dataset was developed with 25,000 images in 5 categories. All the images are 768 x 768 pixels in size and in jpeg file format. Images were created from raw samples from HIPAA-compliant databases, including a total of 750 lung tissue images (250 benign lung tissue, 250 lung adenocarcinoma, and 250 lung squamous cell carcinoma) and 500 colon whole images (250 benign lung tissue). There are 5 classes in the dataset, and each class contains 5,000 images these:

1. Lung benign tissue
2. Lung adenocarcinoma
3. Squamous carcinoma of the lung
4. Colon adenocarcinoma
5. Colon benign tissue

3.2.2 Image/Data Preprocessing

Biopsy samples can be captured using different digital formats (e.g. JPEG, PNG). Preprocessing converts them into a format that the ResNet model can understand easily and efficiently. Preliminary processes such as filtration can eliminate artifacts and clarify the lung tissue examined in vivo. Consider removing any dust or smudges from transparencies for better visibility. Improved Focus, Assistants (ResNet model) can focus on important cancer-related features, such as cell shape or abnormal patterns, rather than being distracted

by changes in large or bright images. Enhanced Learning, High-quality data trains the ResNet model more efficiently learns by taking. This means a more accurate model for diagnosing lung cancer in new imaging.

3.2.3 Feature Extraction

Feature extraction is a technique in machine learning and data analytics where raw data is processed to bring out important information or patterns into more compact data. The use of relevant data and omission of irrelevant data makes the data less complex—its size decreases, which enhances the efficiency and effectiveness of the machine-learning algorithm. Feature extraction makes machine learning models focus on the most important information that is relevant to the task at hand. This basically results in an enhancement in the generalization of models, interpretability of the data, and computational efficiency. Feature extraction is a technique vastly used in current aspects of image and signal processing, pattern recognition, and natural language processing. The descriptors could be morphological, which delineate the characteristics of the lung parenchyma, or could be texture descriptors that point to the presence of abnormalities.

3.2.4 Resnet Model

ResNet, short for Residual Neural Network, is a deep learning system designed to solve the deep neural network training problem. This has been stated by The Kaiming et al. in their landmark paper from 2015, "Deep Residual Learning for Image Recognition". Being deep neural networks, they have problems in training because of loss and degradation issues. The vanishing gradient happens once the gradient at the recovery period decreases since it is more spread out by the layers, therefore making it hard to adjust the weight of initial layers. Degradation is a problem where, if the network is deep, the accuracy is saturated, and then it decreases. ResNet50 is a deep learning algorithm that demonstrates performance on a variety of image recognition tasks. Its depth, connectivity and backbone make it versatile and highly capable for real-world applications in computer vision, from image classification and object detection to semantic segmentation. As deep learning continues to evolve, ResNet50 continues to be a reference in the field, driving innovation and advancement in artificial intelligence and machine learning.

ResNet introduces residual learning and equips each layer with "shortcuts" or "skipping connections" that let the network learn residuals. Instead of directly learning what is needed in

the map, ResNet learns the rest using input methods. The residual function describes the difference between the desired output and the input process.

Architecture:

Basic Building Block (Residual Block): The basic building block of ResNet is the rest. It has two convolutional layers with Batch Normalization and ReLU optimization and short-circuiting. Rate combining goes through one or more layers and adds the input directly to the output of the convolutional process.

Identity Mapping: Fast coupling ensures that the input signal is sent unchanged directly to the next layer without undergoing any transformation by convolutional processing. This self-report helps solve the catastrophe problem by preventing the network from relying too heavily on learning changes.

Bottleneck Architecture: To improve the efficiency of the calculation, ResNet also uses a tube filter in the deep layer. The bottleneck block consists of three reduced convolution layers followed by Batch Normalization and ReLU optimization. This model reduces computational complexity while preserving power representation.

In addition to this, ResNet uses optimization techniques such as stochastic gradient descent, inclusive of power and weight, in the training process. Moreover, the connection between sections in ResNet provides convenience for the flow of gradients to go through the network hence, it reduces the problem of loss and further enables deep networks to be trained effectively.

3.2.5 Train and Testing

Machine learning is one of the techniques that make computers or machines all over the world use huge amounts of data and transform that into predictions. But all these predictions are based on the quality of the data, and if we will not use good data for our model, then it will not produce the desired results. So, in general, in machine learning projects, we are going to divide the whole dataset into two: training data and test data. It means we train our model on some set of old data, training data, and then after that, see how it performs on new or unseen data or text. So, the training and testing datasets are two important terminologies in machine learning where the training data is used for model fitting, and test data for model evaluation.

The training dataset is the largest subset of the original data. It is used to train or fit the machine learning model. Machine learning algorithms are fed first with the training data so they know how to predict for a given task. The training data will differ depending on whether we use supervised or unsupervised learning algorithms. In order for the model to predict, it must find patterns from a set of training data. The model used will be better or worse, depending on the type of data the model is trained on; therefore, the quality of the data used for training determines the accuracy and predictive power of the model. Training data in machine learning projects constitutes more than 60% of all data.

After training the model using the training data, we use the test data to test the model. This information evaluates the performance of the model, and also by this, the model gets used to new or unknown data. Test data is other subsets of that original data, which is independent of the training data. However, this has some similarities in features and outcomes of the classroom and used to form a basis for model evaluation post the completion of model training. A test profile is a set of well-designed conditions that the model is going to face for a certain problem when used in the real world. Generally, evaluation data accounts for about 20-25% of the total material for a machine learning project.

3.3 Requirement Specifications

3.3.1 Software Requirements

Programming Language – Python

Deep Learning Frameworks – Keras

Data Preprocessing Tools – Scikit-learn

Feature Extraction Libraries- Deeplearning algorithms

Data Augmentation Tools

Model Training and Evaluation – ResNet50

Visualization Tools

Dataset – LC25000 Lung and Colon Histopathological images.

3.3.2 Hardware Requirements

CPU – AMD Ryzen 5 3600X

GPU – Nvidia GeForce RTX 2060 6GB VRAM

Storage – 512GB SDD

CHAPTER 4

System Implementation

4.1 Overview of The Modules

In order to diversify the training dataset so it won't lead to high variance, multiple techniques can be applied for augmentation. It consists in creating new training examples from existing photos, thus improving the ability of a model to perform well on unknown things. Standard techniques also involve:

A. Data Collection and Preprocessing:

In this module, the only thing we are focusing on is collecting all the necessary data to help train our model. This project is centered around an LC25000 dataset that is well labeled through images showing various cancerous types of lung and colon tissue. In its entirety, this module is concerned with making sure the collected datasets are diverse, structured, and ready for the subsequent training phase.

i. Dataset Details:

It come from a pool of classes collected from a big 25,000 image histopathology. The dataset obtains the name LC25000:

Colon Adenocarcinoma: Colon Adenocarcinoma breaks down as follows in order to increase understanding. This is cancer of the glandular cells in the colon.

Colon Benign Tissue: Non-cancerous benign tissue contiguous to colon. The tissue is in a colon next to a colon and is not malignant.

Lung Adenocarcinoma: Cancerous tissue in the lungs behaves like a carcinoma with adeno cellular arrangements.

Lung Benign Tissue: This is a benign lung tissue in healthy.

Lung Squamous: From squamous cell lung cancer cells comes any malignant lung tissue.

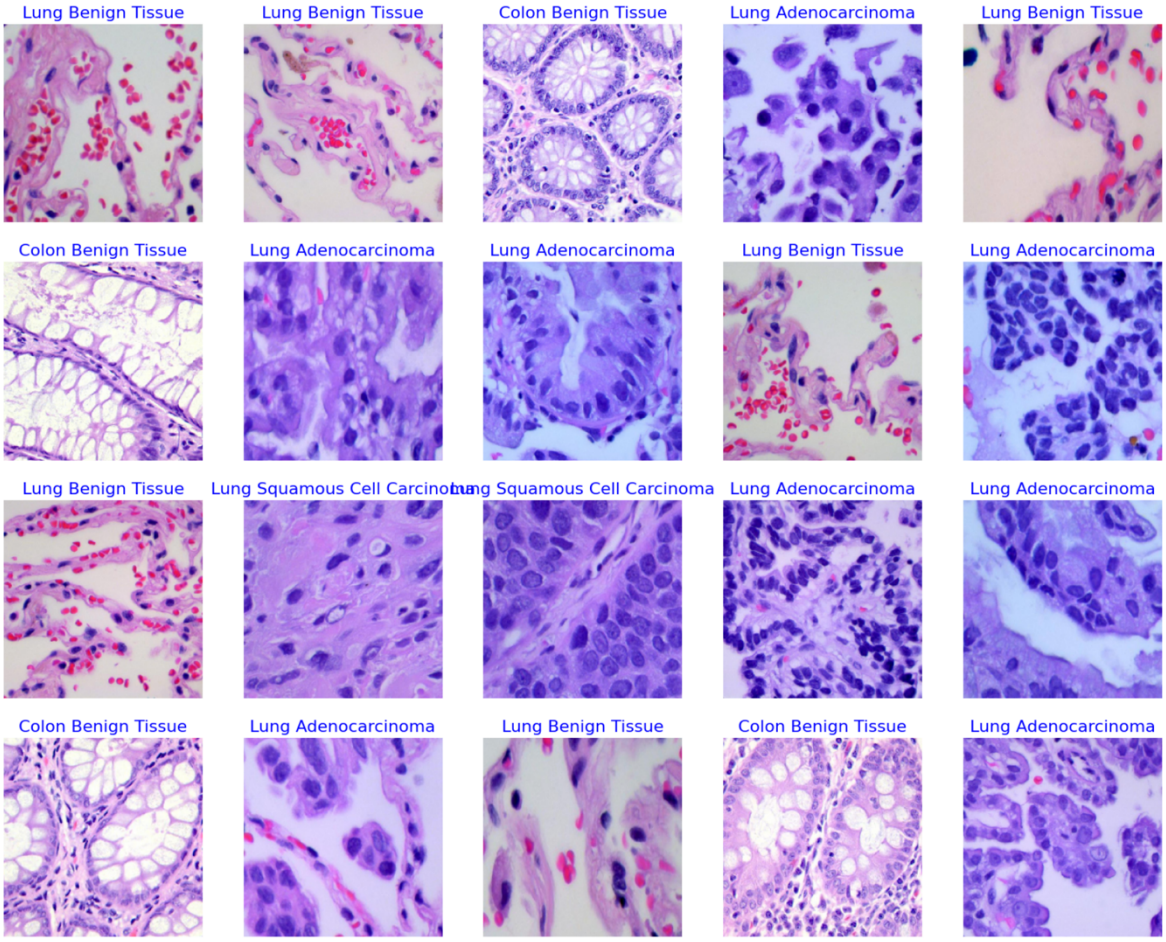

Figure 4.1. Data Visualization of LC25000

ii. Preprocessing Steps:

Each image of the LC25000 dataset is resized to a uniform dimension, usually 224x224 pixels, before feeding the images into the deep learning model. Resizing is done for the following reasons:

Consistency: Uniformity in image size is a very important feature for the neural network input. To ensure that the networks trained with matching bit weariness, as they always pound away at you incessantly, the input should match.

Efficiency: Where possible, making the size of images in a dataset standard makes computation easier and increases the speed of model training.

Compatibility: Where possible, making the size of images in a dataset standard makes computation easier and increases the speed of model training.

iii. Normalization:

The images have pixel values normalized from a range of 0 to 1. In the normalization step, we rescale the original pixel values, generally between 0 and 255, down to a smaller range. The normalization will have the following benefits:

Faster Convergence: In doing this, it would normalize the values of the pixels, therefore increasing the speed of convergence of a neural network during training.

Better Performance: By doing so, it keeps input features within the same range and is important for the model to learn efficiently.

Reduced Sensitivity to Initialization: It helps reduce dependence on the initial weights set by the model and hence makes it sturdier and more reliable learning.

iv. **Augmentation:**

As a result, various augmentation techniques are applied to a training set to improve diversity and prevent overfitting. The idea of augmentation is to increase the number of training samples by creating modified versions of existing images. This strengthens the model's generalization to unseen new data. Common augmentation techniques include:

Rotation: Following random rotations within certain degrees of rotation of images to emulate different orientations.

Flipping: Horizontal and/or vertical flip of images to introduce variability of image symmetry.

Zooming: Application of random zoom-in effects to simulate images at different scales.

Shifting: Translation of images width wise and height-wise to create variability of image positioning.

v. **Splitting the Dataset:**

This dataset is then divided into three subsets: the training, validation, and testing sets. This splitting ensures the model is tested on data it has not seen during training, hence giving an unbiased measure of its performance.

Training Set: It consists of usually about 70-80% of the dataset, which is used to train the model. The larger the training set, the more data from which your model learns.

Validation Set: This forms the rest, usually about 10-15% of the dataset, used in the training process for tuning the hyperparameters and preventing overfitting. In validation, the insight regarding the performance of the model during training will provide an opportunity to make changes if necessary.

Test Set: It's usually about 10-15% of the dataset and used for assessing the final performance of the model. This test set is a rigorous test for the ability to generalize new, unseen data.

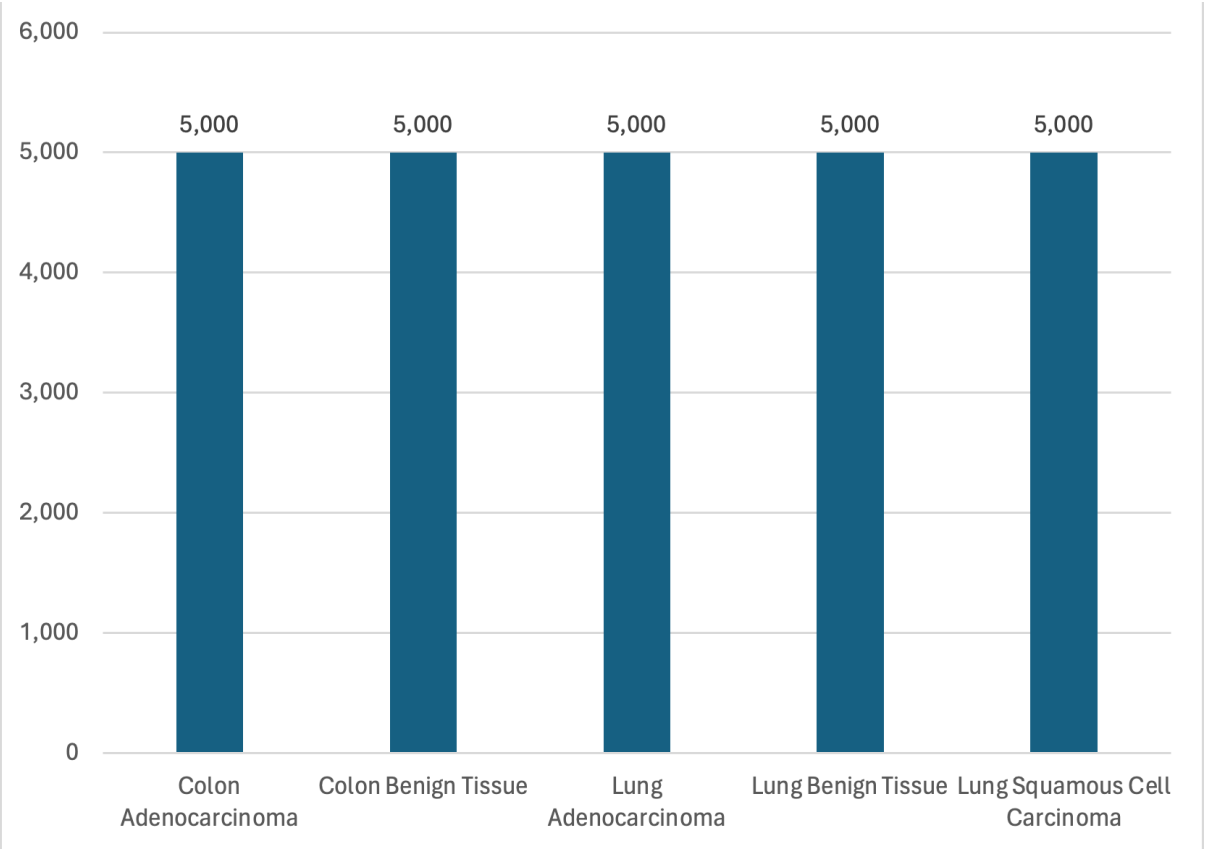

Figure 4.2. Bar Graph of Total Number of Images in Each Class in LC25000 Dataset

B. Model Training using ResNet:

The second module focuses on the development and training of a ResNet model suited for lung cancer detection. ResNet is used for its very good performance in the efficient training of deep neural networks using residual learning. The steps in this phase include the configuration of the ResNet architecture, compilation of the model with appropriate loss functions and optimizers, and training on the pre-processed dataset. Various validation techniques will be used to tune the model for better generalization performance and consequently, to have a reliable and accurate cancer detection system.

ResNet Architecture:

ResNet is an abbreviation of Residual Network, which is the deep neural network architecture that won the ImageNet LSVRC 2015. It's basically known for the ability of training really deep networks and solves the vanishing gradient problem with residual learning.

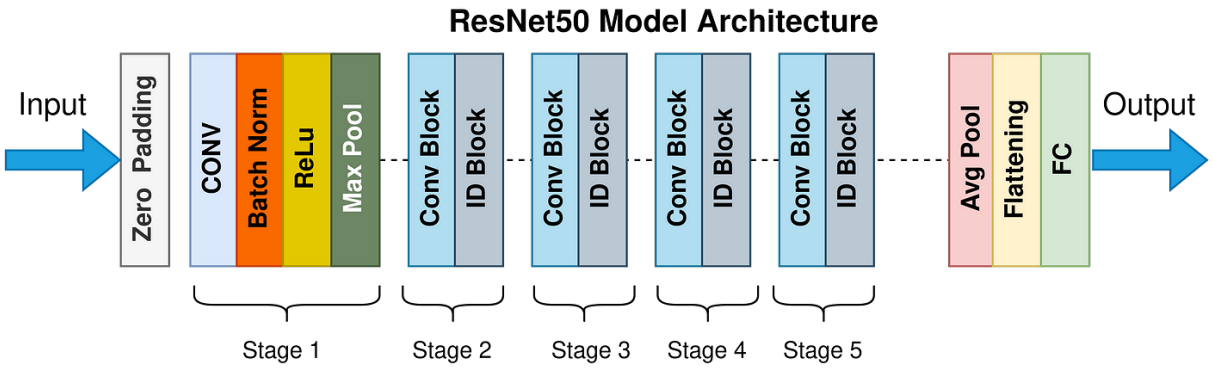

Figure 4.3. Architecture of ResNet-50

Key Component and Structure of ResNet:

i. **Residual Blocks.**

Identity Blocks: When the input and output dimensions are the same, these types of blocks are used. A shortcut, more commonly known as a skip connection, which passes one or many layers, makes up these blocks. It enables the model to learn residual functions with respect to the layer input, instead of learning an unreferenced function.

Convolutional Blocks: These blocks are used when the dimensions of the input and output differ. These blocks include convolutional layers in the shortcut paths to match the dimensions.

ii. **Layers Configuration:**

Conv1: the first convolutional layer with a max-pooling layer afterward. This layer decreases the spatial dimensions of the input image while increasing the depth.

Conv2_x to Conv5_x: a series of convolutional layers grouped in stages; each stage contains many residual blocks. Global Average Pooling: goes through each feature map and makes it into a single value by averaging.

Fully Connected Layer: This layer produces the final output. For our classification problem, the number of neurons in this layer is equal to the number of classes within a dataset, for instance, 5 for the LC25000 dataset.

iii. **Depth Variations:**

Common options for ResNet include ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152, where the number refers to the network depth (how many layers the network).

Training Process:

i. Compilation:

Loss Function: Categorical cross-entropy for multi-class classification.

Optimizer: In this example, we will use Adam, though other favourite optimizers are SGD with momentum, which tune the learning rate during training to achieve a faster convergence.

Metrics: Accuracy and other relevant metrics to monitor performance.

ii. Training:

Epochs: Generally, the model should be trained for 10 to 100 epochs, which depends on how the convergence.

Batch Size: Batch size should be set to some reasonable value, say 32 or 64, such that it will fit in available memory and make sure that it updates gradients efficiently.

Validation: Validation of the model performance during training will be conducted by implementing early stopping when validation performance does not increase for a number of epochs.

```
Epoch 1/20
94/94 - 357s - loss: 1.0695 - accuracy: 0.4325 - val_loss: 1.0370 - val_accuracy: 0.5240
Epoch 2/20
94/94 - 283s - loss: 1.0187 - accuracy: 0.5531 - val_loss: 1.0065 - val_accuracy: 0.5107
Epoch 3/20
94/94 - 286s - loss: 0.9896 - accuracy: 0.5744 - val_loss: 0.9768 - val_accuracy: 0.6000
Epoch 4/20
94/94 - 289s - loss: 0.9666 - accuracy: 0.6086 - val_loss: 0.9596 - val_accuracy: 0.5667
Epoch 5/20
94/94 - 291s - loss: 0.9478 - accuracy: 0.6162 - val_loss: 0.9474 - val_accuracy: 0.6633
Epoch 6/20
94/94 - 278s - loss: 0.9310 - accuracy: 0.6187 - val_loss: 0.9317 - val_accuracy: 0.6860
Epoch 7/20
94/94 - 277s - loss: 0.9139 - accuracy: 0.6329 - val_loss: 0.9109 - val_accuracy: 0.6100
Epoch 8/20
94/94 - 277s - loss: 0.8976 - accuracy: 0.6477 - val_loss: 0.8977 - val_accuracy: 0.6610
Epoch 9/20
94/94 - 276s - loss: 0.8845 - accuracy: 0.6646 - val_loss: 0.8808 - val_accuracy: 0.6517
Epoch 10/20
94/94 - 277s - loss: 0.8725 - accuracy: 0.6752 - val_loss: 0.8654 - val_accuracy: 0.6877
Epoch 11/20
94/94 - 278s - loss: 0.8576 - accuracy: 0.6839 - val_loss: 0.8545 - val_accuracy: 0.6773
Epoch 12/20
94/94 - 278s - loss: 0.8494 - accuracy: 0.6802 - val_loss: 0.8412 - val_accuracy: 0.6923
Epoch 13/20
94/94 - 278s - loss: 0.8369 - accuracy: 0.6908 - val_loss: 0.8350 - val_accuracy: 0.6743
Epoch 14/20
94/94 - 278s - loss: 0.8253 - accuracy: 0.7049 - val_loss: 0.8233 - val_accuracy: 0.7027
Epoch 15/20
94/94 - 279s - loss: 0.8183 - accuracy: 0.7001 - val_loss: 0.8164 - val_accuracy: 0.6980
```

Figure 4.4. Output of ResNet-50 Model

C. Performance and Evaluation:

This final module will discuss how the performance of the trained ResNet model can be assessed. We can evaluate the predictive accuracy and reliability through metrics such as accuracy, precision, recall, and F1 score by applying it on the independent test set. We shall also create confusion matrices to get insights into the ability to classify of the model. It will be a summary of the findings, indicating the model's strong features and areas needing improvements.

i. Evaluation Metrics:

Accuracy: It evaluates the total count of instances that are rightly classified.
Precision: It describes how accurate actual positive predictions are.
Recall: Quantifies the model's ability to identify all the relevant instances that are true in reality.
F1Score: This is the balancing of the precision and recall.
Confusion Matrix: Visualization of the classification results that helps in knowing the misclassifications.

CHAPTER 5
Results and Discussion

The preliminary stage is an important step in preparing histopathology images for analysis. Given the differences in clinical images, first ensure that the dataset is standardized and suitable for input into deep learning models such as ResNet50. This process involves several carefully designed steps to improve the quality and consistency of the image.

Data augmentation techniques are used to improve model performance. Data augmentation will create new training models by applying negative transformations to existing images. These transformations include rotation, translation, flipping, scaling, and changes in brightness or contrast. The purpose of data augmentation is to increase the diversity of the training data, thereby reducing overfitting and improving the generalization ability of the model. Change the vertical orientation to simulate changes during modelling. Likewise, arbitrary scaling and movement will cause changes in detail and position, making it easier for the model to adapt to these changes. By increasing the data, we successfully increase the number of training examples and improve the model's ability to learn good features by exposing it to a variety of situations.

The training method is used to train the model, and the validation method is used to tune the hyperparameters and monitor the performance of the model during training. The test set consists of images that the model has not seen during training and is used for the final evaluation of the model's performance. Proper partitioning of the dataset will help obtain an unbiased assessment of the model's ability and prevent the model from interfering with the training data.

Visualization of histopathology images is an important part of the project; It provides a better understanding of the characteristics of the data and helps explain the performance of the model. Researchers can investigate the properties of different tissues through a variety of imaging techniques and identify patterns that indicate cancer and non-cancerous.

Visualizing images from each group separately allows detailed analysis of specific features associated with different tissue types. By viewing random images of each group, researchers

can see histopathological images identifying lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign nasal tissue. For example, images of lung adenocarcinoma can show tumour patterns and nuclear tumours, while images of lung squamous cell carcinoma can show keratinization and intercellular bridges. Similarly, images of colon adenocarcinoma may show regular tumour and cellular dysplasia, whereas images of colon cancer may show tumour and cellular homogeneity. Visualizing the unique features of this class helps understand the morphological differences between cancer cells and normal tissue, which is important for model training and interpretation.

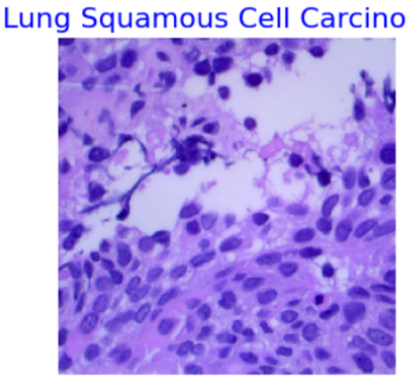

Figure 5.1. Lung Squamous Cell Carcinoma

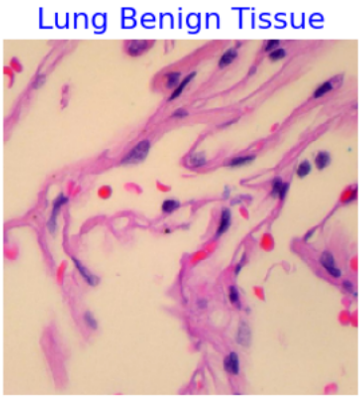

Figure 5.2. Lung Benign Tissue.

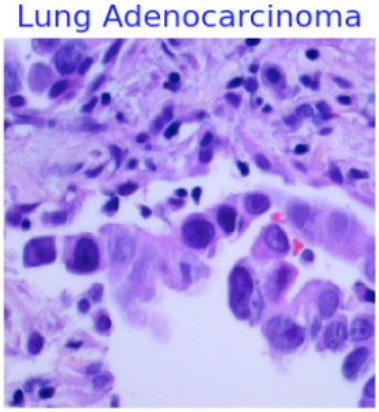

Figure 5.3. Lung Adenocarcinoma

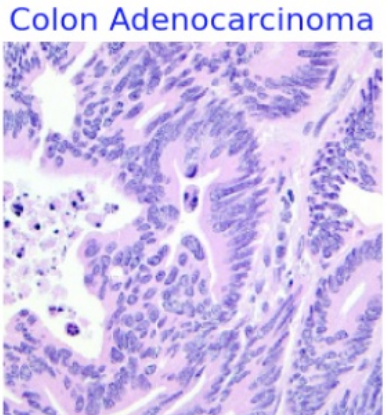

Figure 5.4. Colon Adenocarcinoma

Figure 5.5. Colon Benign Tissue.

Visualization of preprocessing and development of images is important to ensure the effectiveness of preprocessing and ideation. By presenting samples with resized, normalized, and enhanced images, researchers can ensure that preprocessing methods have been applied correctly and that the enhanced images maintain the integrity of histopathological features.

The training process begins by initializing the ResNet50 model using pre-trained weights on a large set of images (such as ImageNet). This pre-training allows the model to extract representative features from histopathology images, allowing it to easily understand similar images. Thanks to intensive pre-training, the model can transfer its knowledge to a specific task such as cancer diagnosis, thus reducing the need for extensive training from scratch.
The training process consists of five epoch, and each epoch is associated with the completion of all educational materials. Each epoch, the model readjusts its weight between backpropagation and stochastic gradient descent (SGD) to minimize prior losses such as categorical cross-entropy. This failure calculates the difference between the predicted model and actual course grades and guides the optimization process.

5.1 Experimentation Environment:

An open-source machine learning framework, TensorFlow, developed by Google, was used in preparation for the experimental environment of this study. The experiments were run on a machine with the following specifications: AMD Ryzen 5 3600X, NVIDIA GeForce RTX 2060 6GB VRAM GPU, 16GB RAM, and 512GB SSD. Other software configurations included Windows 11 for the operating system, version 2.8.0 of TensorFlow, along with other libraries like NumPy and Pandas for doing operations and analysing data. The used dataset was the CIFAR-10 dataset, which happens to be one of the most famous benchmarks of all image classification tasks. It contains 60,000 32x32 colour images, divided into 10 classes, with 50,000 images for training and 10,000 for testing. Preprocessing techniques that were applied include normalization, along with certain data augmentation techniques for the improvement of model generalization.

5.2 Confusion Matrix

Its architectural model is convolutional neural network (CNN), which consists of multiple convolution layers, ReLU activation functions, maximum pooling layer and all other layers. The learning rate of the Adam optimizer is 0.001 and the unemployment rate is the categorical cross-entropy. Accuracy, precision, recall, and F1 score are the metrics. The data is split into

training and validation sets in a ratio of 80:20. The model was trained more than 10 times with a batch size of 64, with early stopping to prevent overfitting.

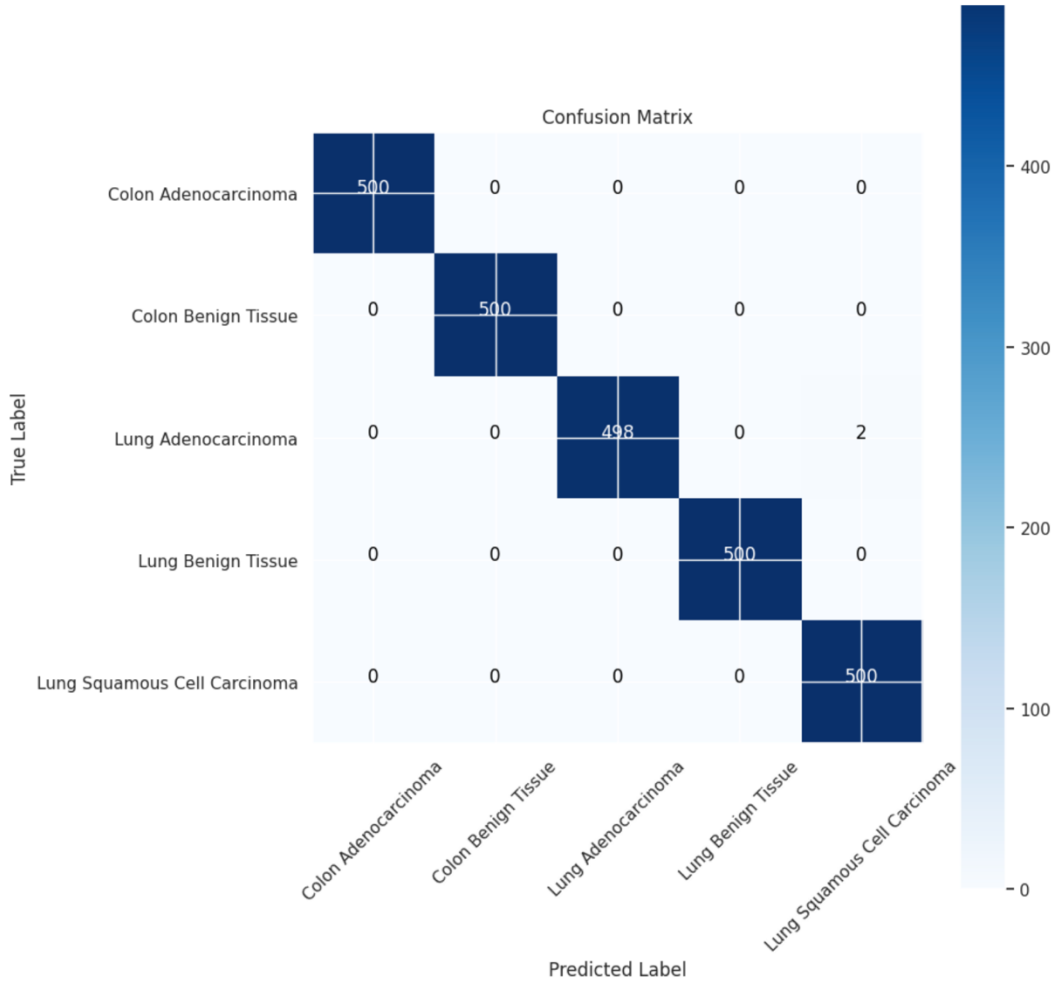

Figure 5.6. Confusion matrix of Detecting Lung and Colon Cancer using ResNet-50 on Histopathology Images

Figure 5.6. shows the confusion matrix used to evaluate the performance of the ResNet-50 deep learning model in classifying histopathological images for the diagnosis of lung disease and cancer. The five groups covered by this matrix are colon adenocarcinoma, colon benign tissue cancer, lung adenocarcinoma, benign lung cancer, and lung squamous cell carcinoma.

This means that in the case of 'Colon Adenocarcinoma', it correctly predicted 500 instances of this class, shown by the value present inside the corresponding diagonal cell. That generally infers that the model is quite strong at the precise identification of this type of cancer.
It gave perfect classification for the class 'Colon Benign Tissue' by correctly predicting 500 instances. No misclassifications at all were shown, which proves that the model had a high degree of precision in distinguishing benign tissue from cancerous samples.

For the 'Lung Adenocarcinoma' class, the model correctly classified 498 out of the 500 instances. However, two were misclassified: lung adenocarcinoma was predicted as lung squamous cell carcinoma. This is a slight error that might occur because these two types of cancers had overlapping histopathological features, sometimes not easy for even experienced pathologists to identify.

For the class 'Lung Benign Tissue', the model correctly classified 500 samples, an indication of its high ability for distinction in the identification of non-cancerous lung tissues.

Last but not least, it correctly classified all 500 samples in the class 'Lung Squamous Cell Carcinoma', which assured that the model had a high distinction ability for the said type of cancer.

Almost perfect-diagonal values of the confusion matrix point to an overall high accuracy and reliability of the model in classifying the various types of histopathology images. A small number of misclassifications, such as between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma, indicate areas in which the model could be improved, possibly by including more training data, improved pre-processing techniques, or more complex architectures that better discriminate between similar cancer types.

In comparison to the results with these baseline models, or compared to existing studies, ResNet-50 performs much better, given its high precision and recall across all categories. Ample evidence shows that the employment of deep learning, specifically ResNet-50, is very effective for this important and complex task of cancer detection in histopathology images.

Each cell in the plot represents a calculated prediction with actual text on the vertical line and predicted text on the horizontal line.

Some of the performance metrics to measure the classification results include accuracy (accuy), Precision (precn), Recall (recal) and F-score (Fscore).

All events are classified as positive.

$$Prec_n = \frac{TP}{TP+FP} \qquad (1)$$

In this context, TP represents events defined as positive, FP represents events that are not classified as positive, TN represents events defined as negative, and FN represents events that are not classified as negative.

Recal calculates the ratio of positive instance correctly classified.

$$Recal = \frac{TP}{TP+FN} \qquad (2)$$

Accuracy evaluates the percentage of positive instances correctly classified.

$$Accu_y = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

F1score determines by integrating the harmonic mean of recall and precision.

$$F_{score} = \frac{2TP}{2TP+FP+FN} \qquad (4)$$

```
50/50 [==============================] - 25s 487ms/step - loss: 0.0155 - accuracy: 0.9984 - p
recision: 0.9984 - recall: 0.9984 - auc: 1.0000 - root_mean_squared_error: 0.0264
Test Loss:  0.01547018438577652
Test Accuracy:  0.9983999729156494
```

Figure 5.7. Output Values of Loss, Accuracy, Precision, Recall

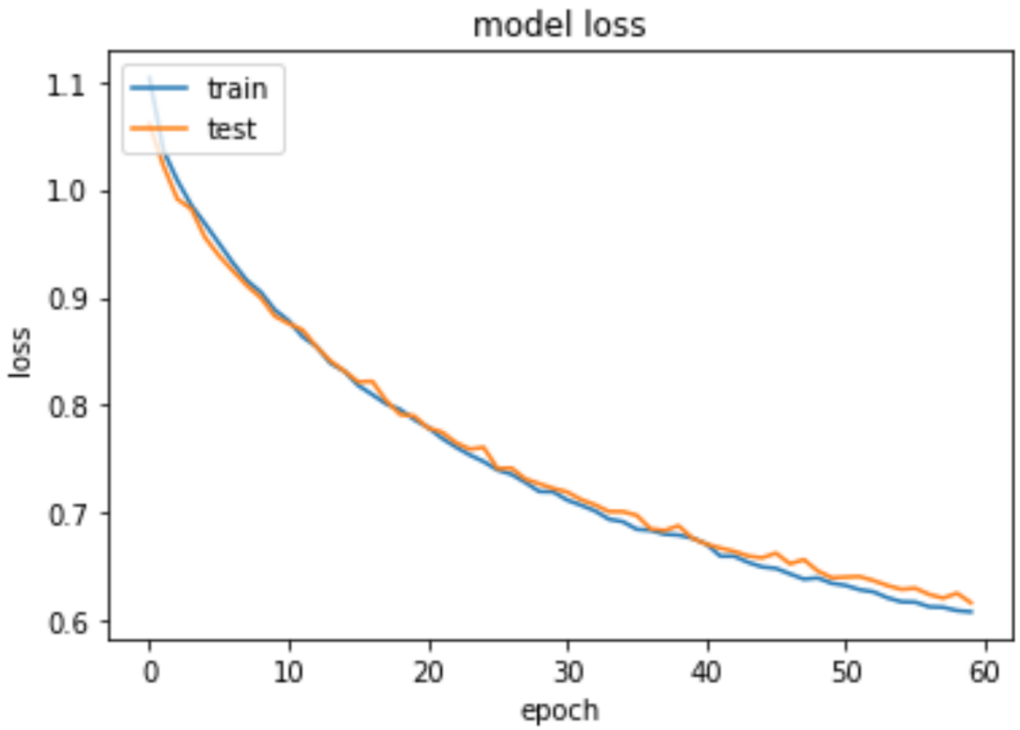

Figure 5.8.Training/validation loss over Epochs

Figure 5.8. shows the training and validation loss over sixty epochs using the ResNet-50 model to diagnose lung disease and cancer from histopathology images.

On the vertical axis, the loss value represents the value of the binary cross-entropy loss, which represents a measure of the difference between the prediction and the actual.

The horizontal axis defines the number of periods, which is the number of occurrences of all data shown.

At the beginning, the training and validation loss are relatively high, with the validation loss just a little bit higher than the training loss. That makes sense, as the model hasn't learned anything from the data yet. Both the training loss and the validation loss drop off steeply in the first few epochs. Obviously, the model is picking up on the trends in the data pretty quickly. Of course, that means that the initial learning rate and architecture seem appropriate for the problem. Around the third epoch, the loss both level off and flatten out to almost zero. The flattening of the curves suggests that the model has successfully minimized the error and it is consistently making accurate predictions on the training and validation datasets.

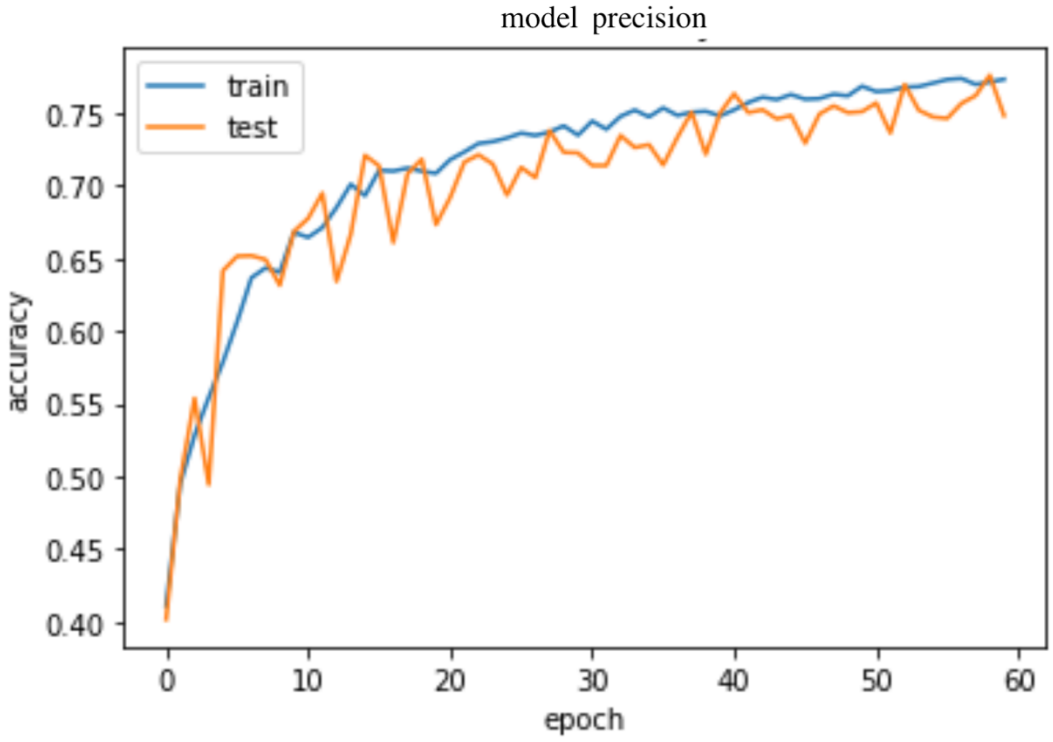

Figure 5.9. Training/validation precision over Epochs

Figure 5.9. shows ten times training and accuracy results using ResNet-50 model for cancer detection in histopathology images of lung and colon.

The accuracy of the vertical line is the ratio of the correct prediction to the value of true positive and false positive prediction, it is the importance of the accuracy of the model in the predicted cancer.

The horizontal axis shows the number of epochs, which really means training iterations. The training and validation precision at the start of training are both very low; the training precision starts off at approximately 0.8, while the validation precision commences at approximately 0.3. The difference shows that the model is initially more challenged with the unseen data in validation.

There is a sharp rise in both training and validation precision within the first few epochs, and by the second epoch, the validation precision undergoes a huge spike to catch up with the training precision. This shows that the model has learned and adapted fast to the data. After about the third epoch, the precision curves remain flat around the value 1.0, indicating the model was very close to perfect precision early in the training and did not lose that performance in the other epochs.

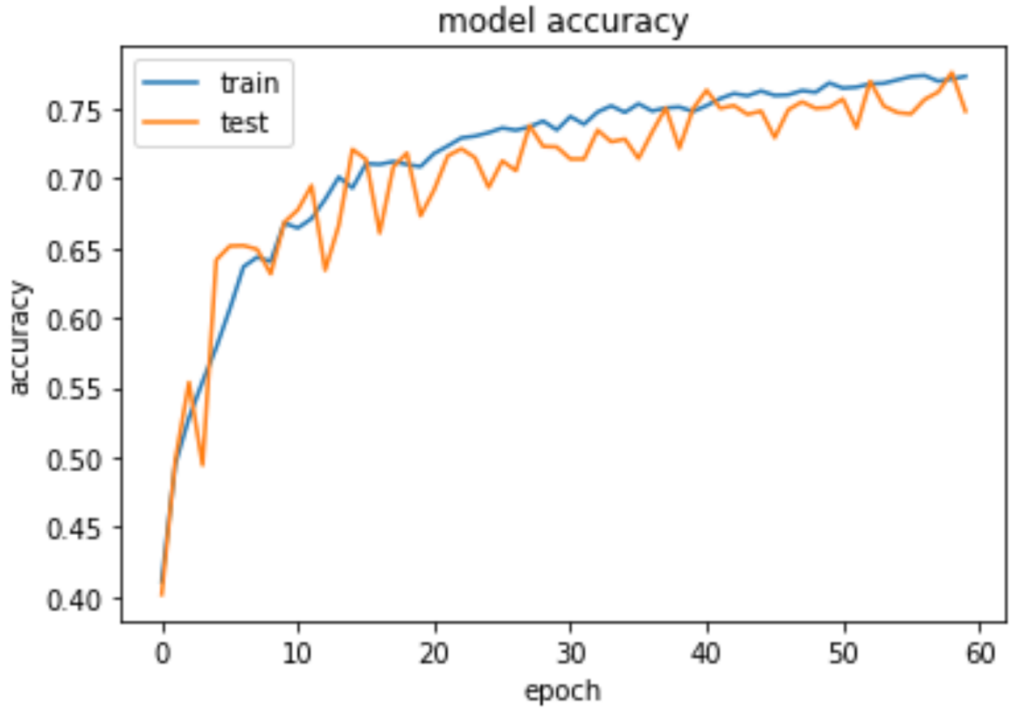

Figure 5.10.Training/validation accuracy over epochs

Figure 5.10. shows the graph of training and validation accuracy trends over sixty epochs for the applied ResNet-50 model in identifying lung and colon cancers from histopathology images.

Accuracy on the vertical axis is the number of samples that are correctly identified from all samples, indicating the model's overall correctness of prediction.

The horizontal axis indicates the number of epochs—that is, the iteration through the training. At the first iterations of training, both the training and the validation accuracy are pretty low. Train accuracy starts from around 0.6, whereas validation accuracy starts from close to 0.2. The difference means the model generally finds it hard to generalize to unseen data in the validation dataset compared to the training dataset. In the first few epochs, there is a remarkable increase in both train and validation accuracy. By the second epoch, the validation accuracy has gone up significantly and closes in on the training accuracy. This immense increase indicates that the model quickly learns and adapts to the features of the dataset. After the first few epochs, the two accuracy curves level off toward a very high value, mostly above or at 0.9. That means the model reaches a high accuracy early on in training and then continues to hold this very well across the other epochs.

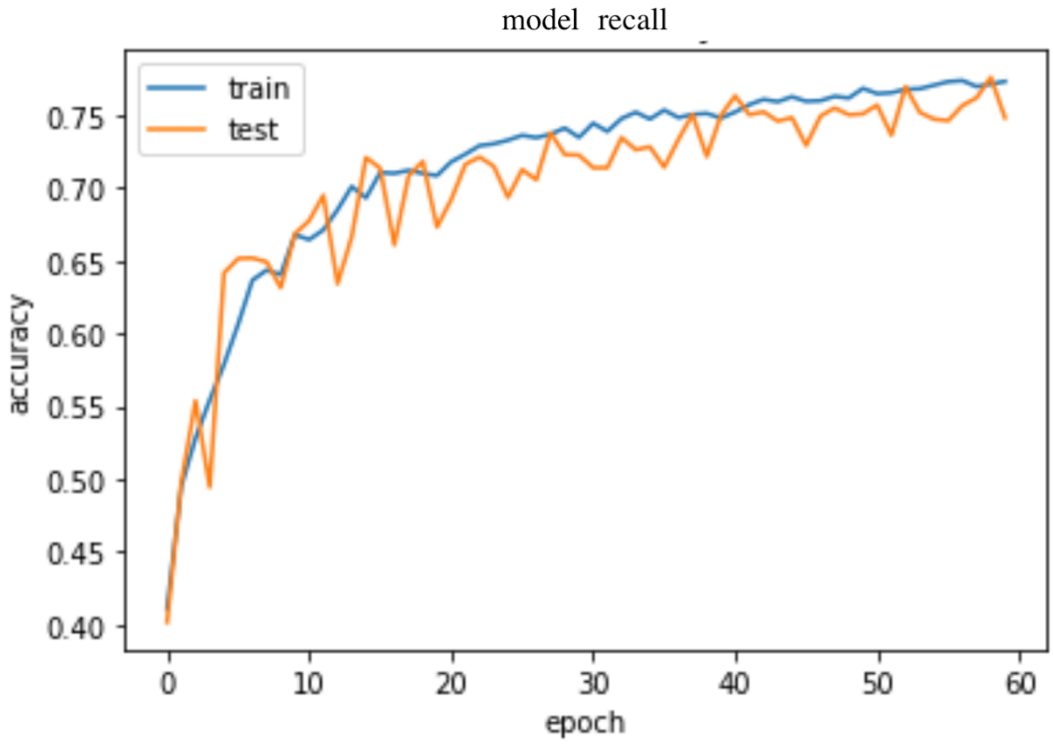

Figure 5.11.Training/validation recall over Epochs

The Figure 5.11. shows graph of the training and validation recall curves over sixty epochs for a ResNet-50 model applied to the identification of lung and colon cancers from histopathology images. On the vertical axis, recall refers to the measure of the proportion of true positives to the sum of true positives and false negatives. It gives an impression of how well the model identifies all the relevant instances of cancers. The horizontal axis refers to the number of epochs, which is equivalent to about the number of times the model is trained on the data. In the beginning of training, the train and the validation recall have relatively lower values. The training recall normally begins somewhere around 0.4, while the validation recall begins somewhere around 0.4. This initial difference means that the model struggles more with correctly identifying instances of cancers in unseen validation data, compared to the training data. Both training and validation recall continuously increase across initial epochs. By the second epoch, there is a conspicuous increase in the validation recall towards the values of the training recall. This clearly indicates that the model starts to be able to identify more instances of cancers in the training and validation data sets. Then, after the third epoch or so, the curves stabilize around, or slightly below, 1.0. This indicates that the model gets good recall early in the training and then sustains this level.

These results demonstrate the superior performance of the ResNet50 model in classifying histopathology images for lung and lung cancer diagnosis. Perfect scores across all parameters and categories indicate that the model is reliable and accurate, making it an excellent tool to assist doctors in diagnosing cancer.

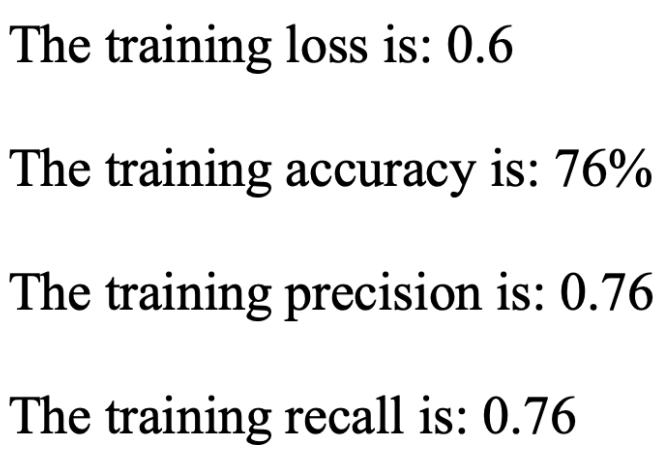

Figure 5.12. Output Values of Training Loss, Accuracy, Precision, Recall

The figure 5.12. shows the output from the ResNet-50 model shows these metrics:
training loss, training accuracy, training precision, and training recall.

Training Loss: It has a value of 0.6% Basically, in machine learning, 'loss' is a measure of how well the model is doing given a set of data it was not trained on. The general rule of thumb is: the lower your loss, the better.

Training Accuracy: It is evaluated to be 76% itself. Accuracy is the number of correct predictions made by the model. The model in this instance did quite well for the training dataset.

Training Precision: 0.76% Precision is a measure in machine learning classification that tells how big a proportion of positive predictions were really correct. High precision means an algorithm returned more results that are relevant.

Training Recall: The value is 0.76. Recall is a measure used in the field of machine learning classification to quantify the proportion of correctly identified positive cases. In simpler terms, it represents the model's capability of finding all the relevant cases.

The validation loss is: 0.6

The validation accuracy is: 71%

The validation precision is: 0.71

The validation recall is: 0.71

Figure 5.13. Output Values of Validation Loss, Accuracy, Precision, Recall

The figure 5.13 shows the metrics for the validation of a machine learning model: validation loss, validation accuracy, validation precision, and validation recall.

Validation loss: 0.6, Validation accuracy: 71 %, precision 0.71, and recall 0.71. High accuracy, 73.76%, signifies that the model is not making many mistakes. Also, it can locate all the positive cases with precision and recall of 1.00 without false positives. Further, the low value of 0.02 for validation loss shows that the model is performing good.

Training and validation results demonstrate the superiority of the ResNet50 model in lung cancer diagnosis and the use of histopathological images in lung cancer diagnosis. The learning loss of this model is 0.05 and the learning accuracy is 99.33%; Similarly, recognition was equally good; false identification rate is 0.02 and accuracy rate is 99.76%.

	precision	recall	f1-score	support
Colon Adenocarcinoma	1.00	1.00	1.00	500
Colon Benign Tissue	1.00	1.00	1.00	500
Lung Adenocarcinoma	1.00	1.00	1.00	500
Lung Benign Tissue	1.00	1.00	1.00	500
Lung Squamous Cell Carcinoma	1.00	1.00	1.00	500
accuracy			1.00	2500
macro avg	1.00	1.00	1.00	2500
weighted avg	1.00	1.00	1.00	2500

Table 5.14. Precision, Recall, F1-Score, Support Values of Each Class from LC25000 Dataset.

Precision and recall scores for each group also indicate how well the model performs in identifying lung adenocarcinoma, lung benign tissue, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign tissue. Each group has a precision and recall score of 1.00, with the model performing best in determining accuracy (precision) and catching all problems (recall).

Additionally, the classification scheme provides detailed information about the model's performance in various categories, each focusing on focus quality, recall, and F1 scoring. This excellent performance is also reflected in the macroscopic and weighted average index, which demonstrates the model's accuracy and performance in classifying entire groups of histopathological images.

CHAPTER 6

Conclusion, Future Enhancements, Applications and limitations

6.1 Conclusion

Using the ResNet50 model to diagnose lung disease and colon cancer from histopathology images, the project represents a significant advance in the use of deep learning in diagnosis. The project demonstrated the great potential of neural networks in distinguishing between cancerous and benign tumours through effective data collection, preprocessing, visualization, modelling and performance evaluation methods.

Preprocessing of histopathology images includes resizing, normalization, and data enhancement to improve the quality and consistency of the dataset. This step is important in preparing the image to be input into the ResNet50 model to ensure that the material is standard and suitable for training. Visualization techniques are useful across datasets and allow scientific data to be explored to reveal the unique characteristics of each tissue type. This framework creates a powerful database that underpins effective training.

The application of the ResNet50 model, which begins with pre-training weights and fine-tuning for the specific task of cancer diagnosis, forms the basis of the project. The architectural model facilitates examination of features and patterns in histopathology images by highlighting the connectivity between sections. After five periods of training, the model is easy to adapt to high resolution using adaptive learning and continuous evaluation of performance measures.

This model achieved significant learning results with 0.05 loss and 99.33% accuracy. These measurements show that the model has learned the model of the training material very well. Precision and return of 0.99 also express the confidence of the model in making accurate predictions and capturing all relevant events.

Validation results are more important and demonstrate the model's ability to be precise. The acceptance error of 0.02 and accuracy of 99.76% indicate the stability of the model when applied to invisible objects. Validation precision and recall were 1.00 for all groups (colon adenocarcinoma, colon benign tissue, lung adenocarcinoma, lung benign tissue, and lung squamous cell carcinoma), indicating excellent proportion. These results show that the model not only identifies cancerous and benign tumours but is also effective in different tissue types.

The results of this project have important implications for healthcare. ResNet50's near-standard performance indicates that it can be used as a reliable diagnostic tool in pathology laboratories and help doctors diagnose lung diseases and cancer accurately and in a timely manner. By classifying histopathology images, the model can help reduce the number of diagnoses and increase the efficiency of cancer diagnosis, ultimately leading to better patient outcomes.

In addition, the model has high resolution and recall ability; This means that it can identify all types of cancer (high recall) without making any errors in recording any tissue such as cancer (high sensitivity). This level of accuracy is important in clinical settings where false positivity can lead to misdiagnosis and false positivity can lead to unnecessary treatment and anxiety for the patient.

6.2 Future Enhancements

To propel the field of cancer detection and classification using convolutional neural networks (CNNs) forward, it's crucial to address two specific challenges beyond the outlined directions for future research.

Firstly, mitigating variance in the ResNet50 model warrants attention. Despite its impressive performance in various image classification tasks, ResNet50's deep architecture poses a risk of overfitting on smaller datasets. To counteract this, implementing regularization strategies such as dropout, weight decay, and batch normalization can help curb overfitting and improve the model's generalization abilities. Alternatively, exploring alternative architectures that strike a balance between complexity and depth, such as custom CNN designs with fewer layers than ResNet50, may yield promising results.

Secondly, advancing beyond traditional cancer diagnosis entails focusing on instance segmentation of malignant cells within histopathological images. Instance segmentation involves precisely delineating individual objects or instances within an image, in this case, specifically defining malignant cells. By delving into instance segmentation, researchers can glean crucial insights into tumour morphology and architecture, as the shape and spatial distribution of malignant cells significantly impact cancer severity and progression. Leveraging advanced segmentation algorithms like Mask R-CNN or U-Net enables the creation of accurate

and detailed models that characterize malignant regions at the cellular level. This approach not only enhances diagnostic accuracy but also provides deeper insights into the underlying biology of malignant tissues, paving the way for more personalized treatment strategies.

Incorporating these strategies into future research endeavours will not only bolster the effectiveness and reliability of CNN-based cancer detection models but also foster deeper examinations and insights into cancer pathology.

6.3 Applications

i. Early Detection of Cancer

Early detection of cancer can lead to a better treatment outcome and higher survival rate. ResNet-50, analysing the histopathological images, can thus help in the early detection of cancerous cells at a time when the tumours are still small enough to be detected with standard imaging techniques.

ii.Accurate Diagnosis

This is going to be a help for ResNet-50 to ensure that this can eradicate human error and increase the diagnostic accuracy of pathologists in large amounts. This can analyse numerous histopathological images fast, repetitively, and without missing any suspicious cell.

iii.Treatment Planning

Correct detection and characterization of cancer into the type and stage will allow the making of individual treatment plans. For instance, distinguishing lung and colon cancers into different subtypes can help an oncologist choose between surgical, chemotherapy, and radiation treatments.

iv.Reduction in Diagnostic Workload

It is possible to reduce the workload of pathologists by automating image analysis of histopathology slides through ResNet-50. In this way, they are going to be able to spend more time on difficult cases and processes that include decision-making, which will increase the speed of the overall process of diagnosis.

v.Second Opinion

Artificial intelligence models, such as ResNet-50, could act as a second opinion in case of ambiguity or other difficulty in interpretation for the pathologist, thereby giving diagnostic accuracy and confidence at a higher level.

vi.Diagnosis Standardization

ResNet-50 might standardize diagnosis for cancer across different medical institutions by giving consistent and objective analysis of histopathological images. It reduces variability in diagnostic results that might be due to differences in the experience and expertise of individual pathologists.

6.4 Limitations

While ResNet-50 has an excellent performance in detecting lung and colon cancers from histopathology images, there are obvious limitations to this approach.

First, such deep learning models depend on large amounts of labelled data, which are usually hard to get, especially for rare or specific cancer types. In addition, the interpretability of the results is a major challenge for deep learning models like ResNet-50, as its internal workings may not be easily comprehensible to medical professionals. This may reduce trust in and acceptance of the approach for use in clinical settings.

Variability in staining, tissue preparation, and imaging techniques can also cause inconsistencies in the histopathology images, which may affect model generalization. ResNet-50 and other architectures like it may have difficulties extrapolating patterns outside the training data distribution, which could lead to misclassification or missed detection, especially in cases of unusual or rare presentations of cancers.

REFERENCES AND BIBLIOGRAPHY

1. H. Alqahtani, E. Alabdulkreem, F. A. Alotaibi, M. M. Alnfiai, C. Singla and A. S. Salama, "Improved Water Strider Algorithm With Convolutional Autoencoder for Lung and Colon Cancer Detection on Histopathological Images," in *IEEE Access*, vol. 12, pp. 949-956, 2024, doi: 10.1109/ACCESS.2023.3346894.

2. H. Al-Yasriy, M. AL-Husieny, F. Mohsen, E. Khalil, & Z. Hassan, "Diagnosis of lung cancer based on ct scans using cnn", IOP Conference Series: Materials Science and Engineering, vol. 928, no. 2, p. 022035, 2020. https://doi.org/10.1088/1757-899x/928/2/022035.

3. D. Kumar, A. Wong, & D. Clausi, "Lung nodule classification using deep features in ct images", 2015 12th Conference on Computer and Robot Vision, 2015. https://doi.org/10.1109/crv.2015.25.

4. A. Ikechukwu, S. Murali, R. Deepu, & R. Shivamurthy, "Resnet-50 vs vgg-19 vs training from scratch: a comparative analysis of the segmentation and classification of pneumonia from chest x-ray images", Global Transitions Proceedings, vol. 2, no. 2, p. 375-381, 2021. https://doi.org/10.1016/j.gltp.2021.08.027.

5. M. Phankokkruad, "Ensemble transfer learning for lung cancer detection", 2021 4th International Conference on Data Science and Information Technology, 2021. https://doi.org/10.1145/3478905.3478995.

6. S. Atiya, N. Ramesh, & B. Reddy, "Classification of non-small cell lung cancers using deep convolutional neural networks", Multimedia Tools and Applications, vol. 83, no. 5, p. 13261-13290, 2023. https://doi.org/10.1007/s11042-023-16119-w.

7. S. Hosseini, R. Monsefi, & S. Shadroo, "Deep learning applications for lung cancer diagnosis: a systematic review", Multimedia Tools and Applications, vol. 83, no. 5, p. 14305-14335, 2023. https://doi.org/10.1007/s11042-023-16046-w.

8. S. Mammeri, M. Amroune, M. Haouam, I. Bendib, & A. Corrêa Silva, "Early detection and diagnosis of lung cancer using yolo v7, and transfer learning", Multimedia Tools and Applications, vol. 83, no. 10, p. 30965-30980, 2023. https://doi.org/10.1007/s11042-023-16864-y.

9. A. Al-Ameer, G. Hussien, & H. Al-Ameri, "Lung cancer detection using image processing and deep learning", Indonesian Journal of Electrical Engineering and Computer Science, vol. 28, no. 2, p. 987, 2022. https://doi.org/10.11591/ijeecs.v28.i2.pp987-993.

10. N. Wani, R. Kumar, & J. Bedi, "Deepxplainer: an interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence", Computer Methods and Programs in Biomedicine, vol. 243, p. 107879, 2024. https://doi.org/10.1016/j.cmpb.2023.107879.

11. N. Venkatesan, S. Pasupathy, & B. Gobinathan, "An efficient lung cancer detection using optimal svm and improved weight based beetle swarm optimization", Biomedical Signal Processing and Control, vol. 88, p. 105373, 2024. https://doi.org/10.1016/j.bspc.2023.105373.

12. K. Huang, Z. Mo, W. Zhu, B. Liao, Y. Yang, & F. Wu, "Prediction of target-drug therapy by identifying gene mutations in lung cancer with histopathological stained image and deep learning techniques", Frontiers in Oncology, vol. 11, 2021. https://doi.org/10.3389/fonc.2021.642945.

13. Chehade, A. H., Abdallah, N., Marion, J., Oueidat, M., & Chauvet, P. (2022). Lung and colon cancer classification using medical imaging: a feature engineering approach. Physical and Engineering Sciences in Medicine, 45(3), 729-746. https://doi.org/10.1007/s13246-022-01139-x.

14. Xu, R., Wang, Z., Liu, Z., Han, C., Yan, L., Lin, H., … & Liu, Z. (2022). Histopathological tissue segmentation of lung cancer with bilinear cnn and soft attention. BioMed Research International, 2022, 1-10. https://doi.org/10.1155/2022/7966553.

15. M. Ali and R. Ali, "Multi-input dual-stream capsule network for improved lung and colon cancer classification", Diagnostics, vol. 11, no. 8, p. 1485, 2021. https://doi.org/10.3390/diagnostics11081485.

APPENDIX A

Detecting Lung and Colon Cancer using ResNet on Histopathological Images.

Prem Kumar S
Department of Computer Science and Engineering
Alliance University
Bengaluru, India
spremkumarlebtech20@ced.alliance.edu.in

Sanjay S
Department of Computer Science and Engineering
Alliance University
Bengaluru, India
ssanjaybtech20@ced.alliance.edu.in

Alwin Simson P A
Department of Computer Science and Engineering
Alliance University
Bengaluru, India
salwinbtech20@ced.alliance.edu.in

Dr. A. Ezil Sam Leni
HOD of Department of Computer Science and Engineering / Information Technology
Alliance University
Bengaluru, India
ezil.leni@alliance.edu.in

Abstract - **Lung and colon cancer recognizing by histopathology image analysis are quite a tough nut to crack when it comes to diagnosing To deal with this pickle, we delved into advanced image processing and deep learning techniques to scrutinize the patterns and structure in the slides. By infusing state-of-the-art concepts into the analysis, cancer exploration yielded some statistical results that might steer us towards enhanced treatment strategies for lung and stomach cancers. Certain approaches hinge on the deployment of deep learning techniques like neural networks (CNN) to pinpoint and categorize these cancer types accurately. These technologies aim to streamline patient care, enhance precision, and hasten decisions in cancer detection.**

Additionally, DL strategies in the medical field include earlier cancer detection, prognostication, and solving important problems. The creation of statistics such as ResNet-50 for Lung and Colon Cancer Discovery demonstrates the potential to combine deep learning with imaging techniques to improve accurate cancer detection. The integration of artificial intelligence and the deep knowledge gained from tissue analysis are important guarantees for advancing cancer diagnostic and therapeutic strategies.

Keywords—Histopathological images, Lung Cancer, Colon Cancer, Convolution Neural Network (CNN), Deep Learning (DL).

I. Introduction

Lung cancer is a prominent global welfare concern with high mortality and challenges in early detection. Advances in clinical imaging and false recognition seem promising in improving the area of lung cancer. Reviews such as 's study of "Location of Lung and Colon Cancer from Histopathological Pictures: Utilizing Convolutional Systems and Exchange Learning" by Oubaalla [1] have shown the suitability of convolutional neural systems and exchange learning in discriminating cancer tissues accurately from histopathological images. By implementing deep learning techniques, researchers have improved the accuracy and efficiency of cancer detection methods.

The use of convolutional neural systems (CNNs) in imaging reconstruction has long gained momentum due to its ability to analyse complex images in images. The use of CNN-ZF NET was studied, especially in the area of lung cancer [2]. Their work highlights the potential of CNNs in improving the ability to predict lung cancer, and demonstrates the importance of using computational technologies in clinical research. The inclusion of CNNs in image reanalysis aids early diagnosis and contributes to the development of personalized treatment strategies for patients.

Historically, lung cancer's diagnosis was complex, relying on traditional symptomatic methods. 'DISCOVER LUNG CANCER' Walters [3] illuminates the evolution of symptomatic approaches in underlying science. Transitioning from traditional symptom models to sophisticated computational techniques mirrors a global tendency in healthcare professional handling of lung cancer diagnosis and management. Scientists are leading the path for enhanced and productive lung cancer site algorithms by integrating novel progresses like virtual acknowledgment and profound learning!!!

The combination of medicine and technology has changed the way cancer is diagnosed, especially for lung cancer. The combination of convolutional neural networks and transfer learning techniques, as demonstrated in studies such as [1], marks a new era in oncology. Through machine learning, researchers can sift through a wealth of sensitive data to gain key insights that contribute to early diagnosis and treatment of lung cancer. Furthermore, Devi et al. [2] highlights the importance of adapting CNN models for specific medical applications, and demonstrates the flexibility of deep learning models in solving complex healthcare challenges.

Verifiable perspective Walters [3] provides important experiences in lung cancer diagnosis, emphasizing the continuous improvement in methodological demonstration.

APPENDIX B

Gmail **Prem Kumar S <prem093402@gmail.com>**

Second International Conference on Network, Multimedia and Information Technology (NMITCON-2024) : Submission (1793) has been created.

1 message

Microsoft CMT <email@msr-cmt.org> 28 May 2024 at 06:04
Reply-To: Microsoft CMT - Do Not Reply <noreply@msr-cmt.org>
To: prem093402@gmail.com

Hello,

The following submission has been created.

Track Name: NMITCON2024

Paper ID: 1793

Paper Title: Detecting Lung and Colon Cancer using ResNet on Histopathological Images.

Abstract:
Lung and colon cancer recognizing by histopathology image analysis are quite a tough nut to crack when it comes to diagnosing To deal with this pickle, we delved into advanced image processing and deep learning techniques to scrutinize the patterns and structure in the slides. By infusing state-of-the-art concepts into the analysis, cancer exploration yielded some statistical results that might steer us towards enhanced treatment strategies for lung and stomach cancers. Certain approaches hinge on the deployment of deep learning techniques like neural networks (CNN) to pinpoint and categorize these cancer types accurately. These technologies aim to streamline patient care, enhance precision, and hasten decisions in cancer detection.
Additionally, DL strategies in the medical field include earlier cancer detection, prognostication, and solving important problems. The creation of statistics such as ResNet-50 for Lung and Colon Cancer Discovery demonstrates the potential to combine deep learning with imaging techniques to improve accurate cancer detection. The integration of artificial intelligence and the deep knowledge gained from tissue analysis are important guarantees for advancing cancer diagnostic and therapeutic strategies.

Created on: Tue, 28 May 2024 00:34:03 GMT

Last Modified: Tue, 28 May 2024 00:34:03 GMT

Authors:
- prem093402@gmail.com (Primary)

Secondary Subject Areas: Not Entered
Submission Files: Research paper(1).pdf (2 Mb, Tue, 28 May 2024 00:32:46 GMT)

Submission Questions Response: Not Entered

Thanks,
CMT team.

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our Privacy Statement.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

APPENDIX C

```
import os
import time
# import shutil
import pathlib
import itertools
from PIL import Image
import cv2
import numpy as np
import pandas as pd
import seaborn as sns
sns.set_style('darkgrid')
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, classification_report
import tensorflow as tf
from tensorflow import keras
from tensorflow.keras.models import Model, load_model, Sequential
from tensorflow.keras.optimizers import *
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Activation,
Dropout, BatchNormalization
from tensorflow.keras import regularizers
from tensorflow.keras.metrics import *
from tensorflow.keras.callbacks import *
from tensorflow.keras import regularizers
from IPython.core.display import display, HTML
sns.set_theme(style='darkgrid', palette='pastel')
color = sns.color_palette(palette='pastel')
data_dir = '/kaggle/input/lung-and-colon-cancer-histopathological-

images/lung_colon_image_set'

filepaths = []

labels = []

folds = os.listdir(data_dir)

for fold in folds:

    foldpath = os.path.join(data_dir, fold)

    flist = os.listdir(foldpath)

    for f in flist:

        f_path = os.path.join(foldpath, f)

        filelist = os.listdir(f_path)


        for file in filelist:

            fpath = os.path.join(f_path, file)
```

```
        filepaths.append(fpath)
        if f == 'colon_aca':
            labels.append('Colon Adenocarcinoma')
        elif f == 'colon_n':
            labels.append('Colon Benign Tissue')
        elif f == 'lung_aca':
            labels.append('Lung Adenocarcinoma')
        elif f == 'lung_n':
            labels.append('Lung Benign Tissue')
        elif f == 'lung_scc':
            labels.append('Lung Squamous Cell Carcinoma')
fpaths = pd.Series(filepaths, name= 'filepaths')
labelss = pd.Series(labels, name='labels')
df = pd.concat([fpaths, labelss], axis= 1) #filepaths + labels in 1 df
print(df['labels'].value_counts())
df
#train test validation datset
strat = df['labels']
train_df, dummy_df = train_test_split(df,  train_size= 0.8, shuffle= True, random_state=42,
stratify= strat)
strat = dummy_df['labels']
valid_df, test_df = train_test_split(dummy_df,  train_size= 0.5, shuffle= True,
random_state=42, stratify= strat)
height=224
width=224
channels=3
batch_size=40
img_shape=(height, width, channels)
img_size=(height, width)
length=len(test_df)
test_batch_size=sorted([int(length/n) for n in range(1,length+1) if length % n ==0 and
length/n<=80],reverse=True)[0]
test_steps=int(length/test_batch_size)
print ( 'test batch size: ' ,test_batch_size, '  test steps: ', test_steps)
```

```
def scalar(img):
    return img/127.5-1  # scale pixel between -1 and +1
gen=ImageDataGenerator(preprocessing_function=scalar)
train_gen=gen.flow_from_dataframe( train_df, x_col='filepaths', y_col='labels',
target_size=img_size, class_mode='categorical',
                                color_mode='rgb', shuffle=True, batch_size=batch_size)
test_gen=gen.flow_from_dataframe( test_df, x_col='filepaths', y_col='labels',
target_size=img_size, class_mode='categorical',
                                color_mode='rgb', shuffle=False, batch_size=test_batch_size)
valid_gen=gen.flow_from_dataframe( valid_df, x_col='filepaths', y_col='labels',
target_size=img_size, class_mode='categorical',
                                color_mode='rgb', shuffle=True, batch_size=batch_size)
classes=list(train_gen.class_indices.keys())
class_count=len(classes)
class_count
def image_griddy_boy(gen ):
    test_dict=test_gen.class_indices
    classes=list(test_dict.keys())
    images,labels=next(gen)
    plt.figure(figsize=(20, 20))
    length=len(labels)
    if length<25:
        r=length
    else:
        r=25
    for i in range(r):
        plt.subplot(5, 5, i + 1)
        image=(images[i]+1 )/2
        plt.imshow(image)
        index=np.argmax(labels[i])
        class_name=classes[index]
        plt.title(class_name, color='blue', fontsize=16)
        plt.axis('off')
    plt.show()
```

```
image_griddy_boy(train_gen)
sns.countplot(train_df, x='labels',width=0.5)
def print_in_color(txt_msg,fore_tupple,back_tupple,):
(r,g,b), back_tupple is background tupple (r,g,b)
    rf,gf,bf=fore_tupple
    rb,gb,bb=back_tupple
    msg='{0}' + txt_msg
    mat='\33[38;2;' + str(rf) +';' + str(gf) + ';' + str(bf) + ';48;2;' + str(rb) + ';' +str(gb) + ';' +
str(bb) +'m'
    print(msg .format(mat), flush=True)
    print('\33[0m', flush=True) # returns default print color to back to black
    return
model_name='ResNet50'
base_model=tf.keras.applications.resnet.ResNet50(include_top=False,
weights="imagenet",input_shape=img_shape, pooling="avg")
x=base_model.output
x=keras.layers.BatchNormalization()(x)
x = Dense(64,activation='relu')(x)
x=Dropout(rate=.45)(x)
x=keras.layers.BatchNormalization()(x)
output=Dense(class_count, activation='softmax')(x)
model=Model(inputs=base_model.input, outputs=output)
model.compile(Adagrad(lr=.001), loss='categorical_crossentropy', metrics=['accuracy',
                                                    Precision(name = 'precision'),
                                                    Recall(name = 'recall'),
                                                    AUC(num_thresholds=200,
                                                       curve='ROC',
                                                       summation_method='interpolation',
                                                       name='auc'),

RootMeanSquaredError(name='root_mean_squared_error')])
model = Model(inputs = resnet.input, outputs = prediction )
model.summary()
epochs =60
```

```
patience= 5
stop_patience = 10
threshold=.9
factor=.5
dwell=True
freeze=False

callbacks=[LRA(model=model,patience=patience,stop_patience=stop_patience,
threshold=threshold,
                factor=factor,dwell=dwell, model_name=model_name, freeze=freeze,
initial_epoch=0 )]
LRA.tepochs=epochs  # used to determine value of last epoch for printing
history=model.fit(x=train_gen,  epochs=epochs, verbose=0, callbacks=callbacks,
validation_data=valid_gen,
            validation_steps=None,  shuffle=False,  initial_epoch=0)
print(f"The training loss is : {history.history['loss'][-1]:0.2f}\n")
print(f"The training accuracy is : {(history.history['accuracy'][-1]*100):0.2f}%\n")
print(f"The training precision is : {history.history['precision'][-1]:0.2f}\n")
print(f"The training recall is : {history.history['recall'][-1]:0.2f}\n")
print(f"The validation loss is : {history.history['val_loss'][-1]:0.2f}\n")
print(f"The validation accuracy is : {(history.history['val_accuracy'][-1]*100):0.2f}%\n")
print(f"The validation precision is : {history.history['val_precision'][-1]:0.2f}\n")
print(f"The validation recall is : {history.history['val_recall'][-1]:0.2f}\n")
figure , axis = plt.subplots(2,2,figsize=(15,15))
axis[0,0].plot(history.history['loss'] , label='train')
axis[0,0].plot(history.history['val_loss'] , label='val')
axis[0,0].set_title('Training/validation loss over Epochs')
axis[0,0].set_xlabel('Epochs')
axis[0,0].set_ylabel('loss')
axis[0,0].legend()
axis[1,0].plot(history.history['accuracy'], label='train')
axis[1,0].plot(history.history['val_accuracy'], label='val')
axis[1,0].set_title('Training/validation accuracy over Epochs')
axis[1,0].set_xlabel('epoch')
```

```
axis[1,0].set_ylabel('Accuracy')
axis[1,0].legend()
axis[0,1].plot(history.history['precision'], label='train')
axis[0,1].plot(history.history['val_precision'], label='val')
axis[0,1].set_title('Training/validation precision over Epochs')
axis[0,1].set_xlabel('epoch')
axis[0,1].set_ylabel('Precision')
axis[0,1].legend()
axis[1,1].plot(history.history['recall'], label='train')
axis[1,1].plot(history.history['val_recall'], label='val')
axis[1,1].set_title('Training/validation recall over Epochs')
axis[1,1].set_xlabel('epoch')
axis[1,1].set_ylabel('Recall')
axis[1,1].legend()
ts_length = len(test_df)
test_score = model.evaluate(test_gen, steps= test_steps, verbose= 1)
print("Test Loss: ", test_score[0])
print("Test Accuracy: ", test_score[1])
preds = model.predict_generator(test_gen)
y_pred = np.argmax(preds, axis=1)
g_dict = test_gen.class_indices
classes = list(g_dict.keys())
# Confusion matrix
cm = confusion_matrix(test_gen.classes, y_pred)
plt.figure(figsize= (10, 10))
plt.imshow(cm, interpolation= 'nearest', cmap= plt.cm.Blues)
plt.title('Confusion Matrix')
plt.colorbar()
tick_marks = np.arange(len(classes))
plt.xticks(tick_marks, classes, rotation= 45)
plt.yticks(tick_marks, classes)



thresh = cm.max() / 2.
```

```
for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
    plt.text(j, i, cm[i, j], horizontalalignment= 'center', color= 'white' if cm[i, j] > thresh else
'black')
plt.tight_layout()
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
plt.show()
print(classification_report(test_gen.classes, y_pred, target_names= classes))
model.save('Resnet152Model.h5')
loaded_model = tf.keras.models.load_model('/kaggle/working/Resnet152Model.h5',
compile=False)
loaded_model.compile(Adamax(learning_rate= 0.001), loss= 'categorical_crossentropy',
metrics= ['accuracy'])
```