Lung and colorectal cancer recognition using CNN

Nguyen Thanh Luan[0009−0002−6355−2286] and Pham Thanh Dat[0009−0006−6193−6883]

FPT University
{nguyenthanhluan102005, feeddepro}@gmail.com

Abstract. Lung cancer and colorectal cancer are among the most common and deadly cancers worldwide, necessitating early and accurate detection for effective treatment. Traditional diagnostic methods, such as histopathological examination, require expert pathologists and are often time-consuming, leading to potential delays in diagnosis. A key challenge in automated cancer classification lies in the complexity of histopathological images, which can exhibit high intra-class variability and inter-class similarities. Additionally, class imbalance in medical datasets can impact model performance, leading to potential misclassifications.
To address these challenges, we develop a deep learning-based classification model using Convolutional Neural Networks (CNNs) to automatically analyze histopathological images of lung and colorectal cancer. The model is trained on a dataset of 25,000 images, covering five distinct tissue categories. Our approach offers key advantages, including automation, reduced dependency on expert pathologists, and the ability to process large-scale datasets with high efficiency. The use of data augmentation and transfer learning further enhances model generalization and robustness. Despite promising performance, challenges such as misclassification in ambiguous cases and sensitivity to dataset quality remain. Future improvements could focus on integrating attention mechanisms and refining preprocessing techniques to further enhance diagnostic accuracy.

1 Introduction

1.1 Project Implementation Context

Cancer is frequently regarded as a disease of the century, characterized by the unbridled proliferation of cells within tissues, forming a nasty tumor. If not diagnosed and treated in time, these cancerous cells can spread to other tissues and organs, a process known as metastasis. Cancer is one of the leading causes of death worldwide. According to the World Health Organization (WHO), lung cancer and colorectal cancer are among the most common types of cancer, with millions of new cases each year. According to IARC, lung cancer is presently the most common type of cancer worldwide, with **2.5 million** new cases, accounting for **12.4% of total cases**, followed by colorectal cancer (**1.9 million**

cases, 9.6%). In Vietnam, lung cancer and colorectal cancer also rank among the most common types, placing a significant burden on the healthcare ***system ('Lung cancer (17.7%), colon cancer (11%)")*** [1, 2]. Early diagnosis plays a pivotal role in increasing survival rates and reducing treatment costs. However, traditional diagnostic methods, such as necropsies and histopathological examinations, still rely heavily on the private assessment of doctors, which can lead to errors and be time-consuming.

Artificial Intelligence (AI) tools are becoming increasingly popular, transforming various industries, including healthcare. AI methods are being utilized to enhance predictive capabilities, diagnosis, and decision-making processes. In medicine, AI has shown great potential in various applications, such as radiology analysis, pathology examination, decision-support tools for organ allocation in transplants, and predictive models for patient outcomes.

1.2 Description of the Problem and Potential Application of the Product

The challenge posed in this project is to develop an AI model capable of classifying histopathological images into five categories: benign lung tissue, lung adenocarcinoma, lung squamous cell carcinoma, colorectal adenocarcinoma, and benign colorectal tissue. In this project, our team has chosen to apply a Convolutional Neural Network (CNN) model.

CNN is a deep learning neural network architecture specifically designed for image processing tasks. It includes convolutional layers that enable the model to automatically extract features from input data without the need for manual feature engineering, unlike traditional methods. CNN has been proven largely effective in medical image classification tasks, such as detecting breast cancer from mammograms, diagnosing diabetic retinopathy from retinal images, and identifying histopathological patterns in cancer research.

The reason for choosing CNN is its capability to recognize complex features in histopathological images, such as cellular texture, tissue distribution, and morphological characteristics of cancer cells. Additionally, CNN can be optimized using pre-trained models on large datasets, improving accuracy and reducing training time. If successfully implemented, this system can be integrated into medical software to build automated diagnostic systems, reducing the workload for healthcare professionals and providing tangible benefits to patients.

1.3 Main Difficulties and Challenges of the Problem

Despite its effectiveness in medical image recognition, this task presents several challenges. First, histopathological images are highly complex, requiring the model to distinguish subtle details. Second, collecting and processing medical image data frequently face issues related to image quality, variations among biopsy samples, and different image formats. Additionally, the model must be optimized to achieve high performance while avoiding overfitting and ensuring generalization across different datasets. Finally, the lack of medical knowledge,

challenges in model development, and time constraints may impact the model's effectiveness. In short, the above difficulties and challenges are like lessons that help us develop, so it can be said that this project is both an opportunity and a challenge.

1.4 Summary of Project Goals and Methods

This project aims to develop an AI model capable of achieving high accuracy in classifying histopathological images. The initial goal is to achieve an accuracy of over 90% and optimize the model for maximum efficiency. The project's outcomes will not only enhance image classification capabilities but also open new research directions for AI applications in medical diagnosis. To achieve this, our approach involves the following key steps: (1) Collecting and preprocessing data to ensure high-quality input images. (2) Developing and training a CNN model using suitable architectures such as ResNet, EfficientNet, or VGG. (3) Evaluating and optimizing the model using techniques like hyperparameter tuning, data augmentation, and overfitting reduction. (4) Testing the model on real-world data to assess accuracy and practical applicability.

Additionally, the project will compare the results of different architectures to determine the most optimal model.

1.5 Summary of Expected Results/Achievements

If successful, the experience and techniques gained from this project can be applied to improve future models, expanding applications to other types of cancer. Continuous research and development efforts can contribute to building more accurate and effective healthcare support systems, easing the workload for medical professionals, and improving diagnostic quality. This project aims to develop an AI model capable of achieving high accuracy in classifying histopathological images. The initial goal is to achieve an accuracy of over 90%.

If successful, the experience and techniques gained from this project can be applied to improve future models, expanding applications to other types of cancer. Continuous research and development efforts can contribute to building more accurate and effective healthcare support systems, easing the workload for medical professionals, and improving diagnostic quality.

1.6 Paper Structure

The remainder of this paper is organized as follows:

- **Section 2: Related Work and Theoretical Background** – Provides an overview of previous studies on cancer recognition using histopathological images and the theoretical foundations of convolutional neural networks (CNNs) in medical image analysis.

- **Section 3: Proposed Methodology** – Describes the model used in this project in detail, including its architecture, training process, and optimization techniques.
- **Section 4: Experimental Results** – Presents the model's performance on the dataset, evaluates its effectiveness, and compares it with other approaches.
- **Section 5: Conclusion and Future Work** – Summarizes the main contributions of the study, discusses its limitations, and suggests directions for future research..

2 Related Works

Currently, the application of AI in healthcare to assist experts in diagnosing and analyzing diseases or potential symptoms is no longer a distant possibility. For example, Machine Learning is widely used in disease prediction and personalized treatment support, such as analyzing data from millions of patients to predict the risk of diabetes. Natural Language Processing (NLP) helps extract insights from electronic medical records (EMRs) and supports communication through chatbots in healthcare. Computer Vision is applied to analyze X-ray images, MRI scans, and other complex medical data, enabling faster and more accurate disease diagnosis. Reinforcement Learning optimizes radiation therapy processes for cancer patients. Generative AI is used to generate simulated images of cancer cells, a technique exemplified by our group's application of CNN in diagnosing lung and colorectal cancer. Numerous studies have demonstrated that Convolutional Neural Networks (CNNs)—a deep learning model applied across various domains such as computer vision and natural language processing—also play a crucial role in classifying different types of cancer.

2.1 Solutions for Lung and Colorectal Cancer Classification in Medical Imaging

Studies Before 2020 Deep learning models, especially Convolutional Neural Networks (CNNs), were widely applied in the medical field before 2020:

LeNet-based approach (LeCun et al., 1998) [3] In this paper, the authors introduced Convolutional Neural Networks (CNNs) and applied them to handwritten digit recognition. Although the study did not directly mention cancer detection using CNNs, the proposed model laid the foundation for image classification, including cancer detection. Additionally, the paper discussed Graph Transformer Networks (GTNs), a global learning approach to optimizing multi-module recognition systems.

- **Advantages:** Simple architecture, easy to implement, foundational for future CNN models.
- **Disadvantages:** Lacks depth to handle complex medical images like histopathological cancer images; not optimized for large and diverse datasets.

AlexNet for medical imaging (Krizhevsky et al., 2012) [4] This paper marked a breakthrough in computer vision by introducing the deep CNN AlexNet. The model won the ILSVRC 2012 competition, significantly reducing classification error rates thanks to its multiple convolutional layers, ReLU activation, Dropout for overfitting reduction, and GPU training acceleration. While the paper focused on ImageNet classification, its methods have influenced various fields, including medical cancer diagnosis.

- **Advantages:** Deep architecture improves feature learning, effective on large datasets, paved the way for medical applications.
- **Disadvantages:** Prone to overfitting on small datasets like medical images, requires high computational resources (GPU).

VGG for histopathology (Simonyan & Zisserman, 2014) [5] This study by the Visual Geometry Group (VGG), University of Oxford, introduced VGGNet, one of the first deep CNN architectures that achieved outstanding results on ImageNet classification. VGGNet improved performance by using smaller 3x3 convolutional filters while deepening the network, enabling better feature learning without drastically increasing parameters. Although originally applied to natural images, VGGNet has significantly influenced medical image analysis, including cancer detection. Many later studies utilized VGG16 and VGG19 for detecting and classifying cancer in medical images such as histopathology slides, CT, MRI, and X-rays.

- **Advantages:** Effective for complex feature learning, suitable for high-resolution medical images, widely adopted.
- **Disadvantages:** Large number of parameters, high computational cost, not optimized for limited medical data.

These methods have become somewhat outdated due to the lack of modern techniques like attention mechanisms and hardware optimization, limiting their applicability to complex medical datasets.

Studies from 2020 to 2023 During this period, CNN applications in medical imaging saw significant improvements:

Swin Transformer for Lung Cancer Cells (Yuru Chen et al., 2022) [6] This study proposed an automated method for lung cancer cell detection and classification using CNNs and Swin Transformer. Microscopic lung cell images were segmented using Mask R-CNN, followed by individual cell extraction. To emphasize target cells, surrounding cells were blurred with Gaussian blur. The Swin Transformer-based classifier outperformed conventional CNNs like ResNet50, achieving 96.16% accuracy, demonstrating its potential for lung cancer diagnosis.

- **Advantages:** Reduced computational cost due to Swin Transformer's hierarchical attention, high accuracy (96.16%).
- **Disadvantages:** Complex to implement on medical devices, dependent on preprocessing (Gaussian blur).

Deep Learning Ensemble 2D CNN for Lung Cancer (Shah et al., 2023) [7] This paper proposed a deep learning approach using 2D CNNs in an ensemble strategy to detect lung nodules with cancerous potential from CT scans. Instead of relying on a single CNN, the researchers combined three different CNN models with varying layers, kernel sizes, and pooling techniques to improve detection accuracy. The study used the LUNA 16 Grand Challenge dataset, achieving 95% accuracy, outperforming baseline methods.

- **Advantages:** High accuracy (95%), leveraging multiple CNNs for improved effectiveness.
- **Disadvantages:** Computationally expensive due to multiple models, may not be feasible on resource-limited devices.

Deep Learning for Colorectal Cancer Detection (Yu et al., 2021) [8] This study employed CNN with transfer learning to detect colorectal cancer in endoscopic images. The model, based on a pre-trained ResNet, was fine-tuned on endoscopic data to achieve high accuracy in classifying polyps and precancerous lesions. Data augmentation techniques were used to overcome limited medical data, reflecting the trend of improving CNNs during this period.

- **Advantages:** High accuracy due to transfer learning, suitable for limited medical datasets.
- **Disadvantages:** Performance depends on initial training data quality, may degrade on heterogeneous datasets.

Recent trends indicate a shift from pure CNNs to hybrid approaches (e.g., Transformers) or optimization techniques (e.g., ensemble, transfer learning), yielding better diagnostic performance.

Studies from 2024 to Present Recent works focus on advanced methods:

BetterNet for Colorectal Cancer (Sengar et al., 2024) [9] This study introduced BetterNet, an efficient CNN architecture combining residual learning and attention mechanisms to improve polyp segmentation accuracy in colorectal cancer endoscopy. Accurate polyp segmentation is crucial for early colorectal cancer diagnosis. BetterNet facilitates precise segmentation, aiding physicians in timely polyp removal and reducing cancer progression risk.

- **Advantages:** Accurate polyp segmentation, integrates attention for focus on important regions.
- **Disadvantages:** Complex training and deployment on resource-limited devices.

ResNet and EfficientNet for Colorectal Cancer (Abhishek et al., 2024) [10] This study applied deep learning models, particularly ResNet and EfficientNet, to colorectal cancer detection. ResNet tackled vanishing gradient issues, enhancing training efficiency, while EfficientNet leveraged compound scaling to balance depth, width, and resolution, achieving high performance with fewer parameters.

- **Advantages:** ResNet addresses gradient issues, EfficientNet optimizes performance with minimal parameters.
- **Disadvantages:** Requires meticulous fine-tuning, may be inefficient on small datasets.

Improved Transformer for Colorectal Cancer (Zhuanping Qin et al., 2024) [11] This study proposed a colorectal cancer recognition algorithm improving Transformer models. The algorithm incorporated skip connections into U-Net, enhancing multi-level feature extraction, combined with Swin Transformer for global information modeling, achieving 95.8% accuracy on the NCT-CRC-HE-100K dataset.

- **Advantages:** High accuracy 95.8%, exploits Transformer's global feature learning.
- **Disadvantages:** Computationally intensive, requires large datasets for effective training.

2.2 Some Studies Used the Same Methodology as the Project

An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification (Shiwen Shen et al., 2018) [12] The authors introduce a hierarchical semantic CNN (HSCNN) model to classify lung nodules' malignancy in CT images.

- **Advantages:** The model provides high interpretability, which is crucial in medical applications. It also enhances feature extraction for malignancy classification.
- **Disadvantages:** The model is computationally expensive and may require large-scale annotated datasets for effective training.

A CNN and Transformer Model for NSCLC Stage Prediction (Lingfei Wang et al.) [13] This study proposes a model combining a 3D Convolutional Neural Network (CNN) and Transformer to predict N-stage and survival rates of non-small cell lung cancer (NSCLC) patients based on CT image data.

- **Advantages:** The integration of CNN and Transformer enhances feature representation, leading to improved prediction accuracy.
- **Disadvantages:** Transformers require significant computational resources and may suffer from data overfitting when trained on small datasets.

Lung and Colon Cancer Detection Using a Deep AI Model (Nazmul Shahadat et al.) [14] This research introduces a novel deep learning model for efficient detection of lung and colon cancers. The study proposes a lightweight, parameter-efficient model suitable for mobile devices, utilizing a 1D CNN enhanced with Squeeze-and-Excitation layers. The model achieved 100% accuracy on a large dataset of histopathological images, indicating its potential to revolutionize medical diagnostics by enabling faster, more accessible, and reliable cancer screening.

- **Advantages:** The model is lightweight and suitable for deployment on mobile devices, making cancer screening more accessible.
- **Disadvantages:** Achieving 100% accuracy may indicate overfitting, and real-world performance needs further validation.

2.3 Some Studies Used the Same Method but in Different Fields

Traffic Sign Recognition with Multi-Scale Convolutional Networks (Pierre Sermanet and Yann LeCun) [15] This paper describes the application of multi-scale CNNs for traffic sign recognition, an essential component for autonomous driving. The model employs convolutional layers to extract features such as shape, color, and symbols from traffic sign images, classifying them into categories like "Stop" or "Speed Limit." The results demonstrate that CNNs outperform traditional methods like SVM in real-world conditions with varying lighting and viewpoints.

- **Advantages:** CNNs offer superior accuracy in recognizing traffic signs compared to traditional machine learning models. The model adapts well to variations in lighting and viewpoint.
- **Disadvantages:** Requires large amounts of labeled training data, and real-world conditions like occlusion and motion blur may degrade performance.

Using Deep Learning for Image-Based Plant Disease Detection (Sharada P. Mohanty et al.) [16] This study explores the use of CNNs to detect plant diseases through leaf images. Models such as AlexNet and GoogLeNet were trained on the PlantVillage dataset, containing thousands of images of healthy and diseased leaves. CNNs identified disease symptoms (e.g., spots, mildew) with high accuracy, assisting farmers in early diagnosis and effective crop management.

- **Advantages:** The method provides an automated, scalable solution for plant disease detection, reducing reliance on expert knowledge.
- **Disadvantages:** Performance may decline when applied to real-world scenarios with varied environmental conditions, such as lighting changes and different leaf backgrounds.

Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparison Analysis (Mohamed Amine Ferrag et al.) [17] This paper presents the application of CNNs for detecting cyber-attacks in network security. CNN models were trained on network datasets such as NSL-KDD and CICIDS2017 to classify network traffic as normal or anomalous (attacks). CNNs extract spatial-temporal features from packet data, effectively recognizing attack patterns such as DDoS, intrusions, or malware with high accuracy. The study compares CNN performance with other deep learning methods like RNNs and traditional machine learning approaches, showing CNNs' advantages in handling complex network data.

- **Advantages:** CNNs provide a high-accuracy approach to intrusion detection, outperforming traditional methods in identifying cyber threats.

– **Disadvantages:** Training on outdated datasets may lead to reduced effectiveness against evolving cyber threats. Additionally, real-time intrusion detection requires significant computational resources.

2.4 Advantages and disadvantages of CNN

Criteria	**Advantages**	**Disadvantages**
Feature extraction from raw data	CNN can learn from raw pixel data without the need for manual feature extraction, reducing dependence on preprocessing techniques.	CNN requires a large amount of labeled data for effective training, making data collection and preprocessing costly and time-consuming.
Speed and computational efficiency	CNN can automatically extract the most relevant features from images, making training and inference faster and reducing the complexity of input data.	CNN has a large number of parameters to update during training, leading to high computational time, especially when working with large datasets.
Spatial structure exploitation and hierarchical learning	CNN utilizes the spatial structure of data using vector filters while preserving pixel context, enabling recognition of variations and image diversity.	CNN can suffer from **overfitting**, meaning the model memorizes training data details too well, reducing its generalization ability to new data. Techniques such as Dropout, Batch Normalization, and Data Augmentation are needed to mitigate this issue.
Model compression and optimization	CNN can replace fully connected layers with convolutional blocks, optimizing processing speed and reducing the number of parameters in the model.	CNN is often considered a **"black box"**, making it difficult to interpret and explain its decision-making process, posing challenges in fields like healthcare, security, and law.
Performance with limited training data	Since CNN has **strategies for 2D data**, it can learn with fewer examples compared to some Transformer models designed for computer vision.	If the training dataset is too small, CNN may fail to learn effective patterns and perform poorly on new data.
Wide range of applications	CNN is used in various fields such as computer vision, speech processing, automation, security, object recognition, and natural language processing.	CNN requires significant computational resources (GPU/TPU), increasing operational costs and training time.

Table 1. Advantages and Disadvantages of CNN

2.5 Research related to the dataset LC25000

This dataset contains 25,000 histopathological images with 5 classes, central to our project, is a benchmark for cancer classification. All images are 768 x 768 pixels in size and are in jpeg file format.

The images were generated from an original sample of HIPAA compliant and validated sources, consisting of 750 total images of lung tissue (250 benign lung tissue, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas)

and 500 total images of colon tissue (250 benign colon tissue and 250 colon adenocarcinomas) and augmented to 25,000 using the Augmentor package.

Lung and Colon Cancer Histopathological Image Dataset (LC25000) by (Andrew A. Borkowski et al., 2019) [18] This study aims to create a histopathology image dataset to aid researchers in the field of machine learning and artificial intelligence, especially in the diagnosis of lung and colon cancer. The LC25000 dataset consists of 25,000 color images, divided into 5,000 images for each of the five data types. All images are 768 x 768 pixels in size and saved in JPEG format. The original study used a simple CNN to classify 5 classes in LC25000, achieving high accuracy with data augmentation. The common feature is focusing on the efficiency of large datasets.

Classification Of Lung and Colon Cancer Histopathological Images Using Convolutional Neural Network (CNN) Method An A Pre-Trained Models by (Brilly Lutfan Qasthari et al., 2023) [19] The paper focuses on the classification of histopathological images of lung and colon cancer. The main objective of the study is to build a cancer diagnosis support system by automatically classifying histopathological images into five different categories, which will help reduce the burden on medical professionals. The study uses a dataset of 25,000 images, of which 80% are used for training and 20% for testing the model. The images are classified into groups: lung benign tissue, lung adenocarcinoma, lung squamous cell carcinoma, colon adenocarcinoma, and colon benign tissue. The results show that the model achieves an accuracy of 99.96%, demonstrating the effectiveness of using a pre-trained CNN model in cancer classification based on histopathological images.

An Evolutionary Attention-Based Network for Medical Image Classification by (Hengde Zhu et al., 2023) [20] The study focuses on improving the generalizability of deep learning models in the medical field, making them effective on many different datasets instead of just optimizing for a specific disease. The model uses a densely connected attention network (DCA-Net) to automatically assign weights to important features, combined with two evolutionary mechanisms: internal evolution (optimization during training) and inter-network evolution (information exchange between two versions of the model). When evaluated on four public medical datasets, EDCA-Net shows superior performance compared to existing methods on three datasets and achieves comparable performance on the remaining set. This research makes an important contribution to the application of AI in medicine, especially in imaging diagnosis, helping to improve the accuracy and usability of deep learning models in clinical practice.

3 Methodology

3.1 Overview

In this project, we decided to use a Convolutional Neural Network (CNN) model to detect lung and colorectal cancer based on histopathological images. CNN was chosen due to its strong feature extraction capabilities and efficiency in handling image datasets. Initially, the input histopathological images are preprocessed, appropriately labeled, and stored in a dataframe. The dataset is then divided into different subsets for model training. Before being fed into the model, the input images are resized to a fixed size.

Our model consists of 22 layers and 6 main blocks, focusing on layers such as convolutional layers, max pooling layers, fully connected layers, etc. The model is optimized to enhance feature recognition in histopathological images.

The dataset consists of 25,000 histopathological images across five classes, augmented using the Augmentator library in Python from an initial set of 750 images: Colon Adenocarcinoma, Colon Benign Tissue, Lung Adenocarcinoma, Lung Benign Tissue, Lung Squamous Cell Carcinoma

3.2 Data Preprocessing

Reading Data and Creating Dataframe The program uses the `os` library to scan all dataset subfolders and store them in a list, while labels are stored separately. Using the `concat` function from pandas, we create a dataframe from these lists. The dataframe ensures that each image is properly labeled, facilitating easy retrieval of image paths and labels during CNN training.

Splitting the Dataset The dataset is divided into three parts: Training set (80% of the data) is used for model training, Validation set (10% of the data) is used for hyperparameter tuning, and Test set (10% of the data) is used to evaluate the model on unseen images. This split helps reduce bias and prevent overfitting, as a dedicated set is reserved for evaluation.

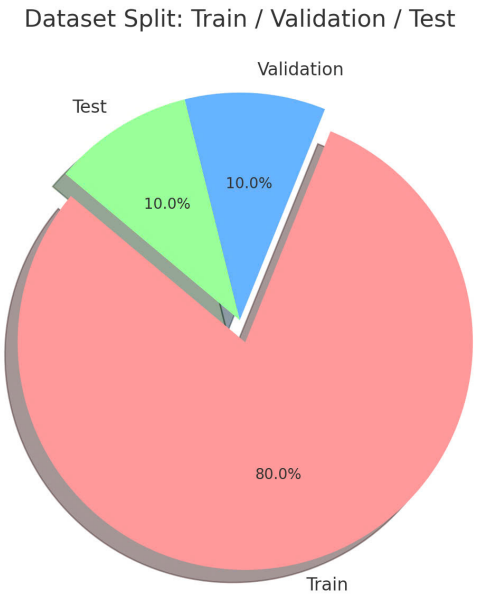


Fig. 1. Dataset Split

Creating Image Data Generator with Data Augmentation In this step, we define hyperparameters and resize images to a fixed size. With a batch size of 64, images are resized to 224x224 pixels with three RGB channels. This is a standard size for many CNN models, ensuring sufficient detail retention for cancer detection.

An Image Data Generator is created to augment the dataset, improving model generalization. Data augmentation includes random rotations, flipping, zooming, and brightness adjustments. These transformations are controlled using parameters in the `ImageDataGenerator` function, such as `rotation_range`, `width_shift_range`, `height_shift_range`, `horizontal_flip`, `zoom_range`, and `brightness_range`. Augmentation is applied only to the training set to help the model learn from diverse variations.

3.3 CNN Model

Overview

Data Formatting Input data is resized to 224x224 pixels. The CNN model consists of 22 layers and 6 main blocks.

Convolutional Blocks

- Block 1: 2 Convolutional layers + 1 Max Pooling (2×2)
- Block 2: 2 Convolutional layers + 1 Max Pooling (2×2)
- Block 3: 3 Convolutional layers + 1 Max Pooling (2×2)
- Block 4: 3 Convolutional layers + 1 Max Pooling (2×2)
- Block 5: 3 Convolutional layers + 1 Max Pooling (2×2)
- Block 6: Fully Connected Layers (FC)

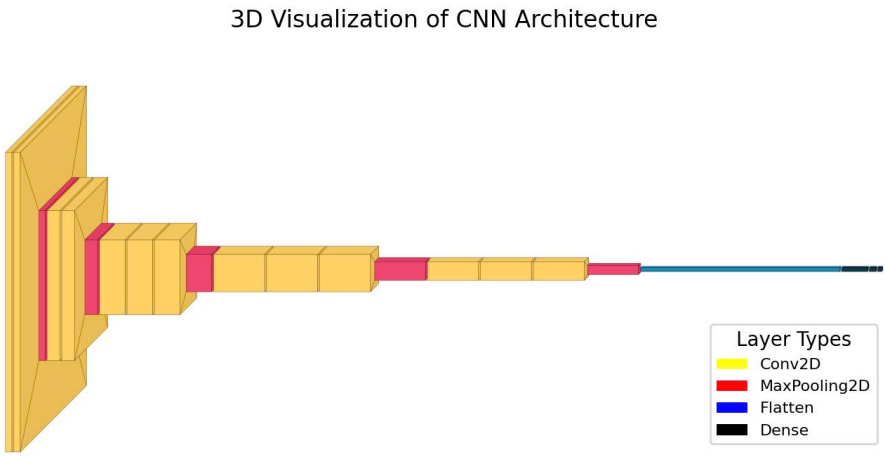


Fig. 2. Structure of model

Layer Analysis and Functions

Convolutional Layers Convolutional layers extract spatial and local features, detecting edges, contours, and cellular structures. Each filter has a size of 3x3, with the number of filters doubling in each subsequent layer, starting from 64 filters. Filters are designed to capture different patterns, such as horizontal edges, vertical edges, etc. Padding is set to "same" to maintain image size.

Given an input image with dimensions (H_{in}, W_{in}, C_{in}), and a filter of size (K, K, C_{in}), the output size is computed as:

$$H_{out} = \frac{H_{in} + 2P - K}{S} + 1 \tag{1}$$

$$W_{out} = \frac{W_{in} + 2P - K}{S} + 1 \tag{2}$$

$$C_{out} = C_{filters} \tag{3}$$

where P is padding, S is stride, and $C_{filters}$ is the number of filters.

Activation Function (ReLU) The activation function plays an important role in helping the model learn complex details with a nonlinear mathematical function, with a linear function calculating simpler details. In each convolutional layer, we decided to use the Relu function as the activation function. ReLu (Rectified Linear Unit) is one of the most popular functions in deep learning models, especially CNNs. The advantage of the ReLu function is its simplicity and it has been proven to speed up the training process. Next, it is not bounded like the sigmoid or Tanh function, so it is not the cause of vanishing gradients.

$$f(x) = \max(0, x) \tag{4}$$

Max Pooling Layers In our CNN model, 2D Max Pooling layers will appear after each convolutional layer. They play an important role in reducing the size of the input image, helping to reduce the number of parameters and increase computational efficiency. In this layer, Max Pooling will reduce the data dimensionality by taking the largest value in the 2x2 size, so the input images will be reduced by half but still retain the features. Such a size reduction will help the next layers not to have to process a large amount of information, reduce overfitting when Pooling helps extract the most important features, make the model generalize better and help speed up training and reduce the need for GPU/CPU memory.

$$O(i, j) = \max(X_{i,j}, X_{i,j+1}, X_{i+1,j}, X_{i+1,j+1}) \tag{5}$$

where X is the input matrix (feature map), and $O(i, j)$ is the value at position (i, j) in the output matrix.

With the input images being reduced in dimension, it will be calculated as follows:

$$H_{out} = \frac{H_{in}}{2}, \quad W_{out} = \frac{W_{in}}{2} \tag{6}$$

Fully Connected Layer (FC) Fully Connected Layer is the layer that connects all neurons between the two layers in the neural network, playing an important role in synthesizing information and making the final prediction. After the image goes through the Convolutional and Pooling layers, it is no longer an image but a set of important features. However, these features are still just numeric values that cannot be used directly for classification. The Fully Connected Layer takes on the task of transforming them into a form that can be used to make prediction decisions. In our model, after going through the Conv2D and MaxPooling2D layers, the output is a multidimensional tensor containing information extracted from the image. The Flatten() layer will convert this tensor into a 1D vector, so that the data can be fed into the Fully Connected Layer. The model includes three Dense layers:

- Dense(256, activation="relu")
- Dense(64, activation="relu")
- Dense(class_count, activation="softmax")

Thus, the Fully Connected Layer helps synthesize information from extracted features, turning them into specific and accurate predictions, which plays an important role in deciding how accurate the image recognition model is.

The FC layer follows:

$$y = f(Wx + b) \tag{7}$$

where x is the input vector, W is the weight matrix, b is the bias, and f is the activation function.

Softmax Activation Softmax Activation is an activation function commonly used in multi-class classification models. It helps transform the model's output into a probability distribution, ensuring that the sum of the probabilities of all classes equals 1. This allows the model to make a decision by selecting the class with the highest probability.

In our model, Softmax is applied in the last layer (output layer) to convert the output from the Fully Connected Layer into a set of probabilities corresponding to the classification labels.

The Softmax function is defined as follows:

$$P(y = c|x) = \frac{e^{z_c}}{\sum_{n=1}^{C} e^{z_n}} \tag{8}$$

where z_c represents the output for class c, and C is the total number of classes.

Our model utilizes `Dense(class_count, activation="softmax")`, meaning that Softmax normalizes the output of the last Dense layer into probabilities for each disease (or non-disease) class. The class with the highest probability is selected as the predicted label.

Loss Function In our model, the loss function used is Cross-Entropy Loss, one of the most popular loss functions for multi-class classification problems. Cross-Entropy measures the difference between the predicted probability distribution of the model and the actual distribution of the labels.

The general formula of Cross-Entropy Loss for a single sample with actual label y and predicted probability $\hat{y}$ is:

$$L = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{9}$$

where y_c is the true label (one-hot encoded) and $\hat{y}_c$ is the predicted probability.

This ensures optimal classification accuracy for histopathological images.

4 Experiment

The purpose of this experimental section is to demonstrate that our CNN model is capable of effectively classifying histopathological images of lung and colorectal cancer, and to provide experimental evidence to draw reasonable conclusions

about the performance and application potential of the method. We have conducted a full range of steps from data preprocessing, model training, validation, to testing, with the goal of comprehensively evaluating the ability to classify 5 different image classes. The obtained results not only demonstrate the accuracy but also clarify the difficulties in the process of building the model.

4.1 Experimental description

The experiment was designed to evaluate the effectiveness of the CNN model in classifying five histopathological image classes, including Colon Adenocarcinoma, Colon Benign Tissue, Lung Adenocarcinoma, Lung Benign Tissue, and Lung Squamous Cell Carcinoma. This evaluation was conducted through the stages of preprocessing, training, validation, and testing. The dataset was collected from [21], consisting of 25,000 images resized to 224x224x3. The data was split into training (80%), validation (10%), and testing (10%) sets. To analyze the data distribution, we used Python to statistically count the number of samples per class, revealing that each class contains approximately 5,000 samples, which is fairly balanced. This distribution was illustrated using a bar chart with the Seaborn library.

Image type	**Total images**	**Training set**	**Testing set**	**Validation set**
Lung Adenocarcinoma	5000	4000	500	500
Lung Benign Tissue	5000	4000	500	500
Lung Squamous Cell Carcinoma	5000	4000	500	500
Colon Adenocarcinoma	5000	4000	500	500
Colon Benign Tissue	5000	4000	500	500

Table 2. Description of the employed dataset

The model was built with 13 Conv2D layers (with filters increasing from 64 to 512), 5 MaxPooling2D layers, 3 Dense layers, and 1 Fully Connected layer. The model utilized hyperparameters such as a learning rate of 0.001 (Adamax), 20 epochs, and a batch size of 64. The training, validation, and testing processes were conducted sequentially to ensure a comprehensive evaluation of the proposed CNN model.

During the training phase, the training dataset (accounting for 80% of the total 25,000 samples) was augmented using TensorFlow's ImageDataGenerator with transformations such as rotation and horizontal flipping to enhance generalization. The model was then trained for 20 epochs with a batch size of 64, using the Adamax optimizer (learning rate of 0.001) and the Categorical Crossentropy loss function.

The validation phase utilized the validation dataset (10%) to monitor performance after each epoch, recording loss and accuracy to detect overfitting. The final results achieved a validation loss of 0.0103 and an accuracy of 99.72

Finally, the testing phase was conducted on the test dataset (10%) to evaluate the model's generalization ability using unseen data. This resulted in a test loss of 0.0131 and an accuracy of 99.56

The entire process was carried out in the Pycharm environment, running on a Ryzen 7 7750HS CPU, 16GB DDR5 4800 RAM, and an NVIDIA GeForce RTX 4050 6GB GDDR6 GPU.

4.2 Describe the model evaluation metrics

To comprehensively evaluate the performance of the CNN model in classifying 5 histopathological image classes (Colon Adenocarcinoma, Colon Benign Tissue, Lung Adenocarcinoma, Lung Benign Tissue, and Lung Squamous Cell Carcinoma), we selected and applied a set of standard metrics, each designed to reflect a specific aspect of model performance, from overall accuracy to fine-grained classification ability on each class. First, Accuracy is used as the main metric to evaluate the overall performance. This metric is chosen because our data is nearly balanced (about 5,000 samples per class), which helps Accuracy accurately reflect the classification ability without being biased by class imbalance. The actual results show that Accuracy reaches 99.56% on the test set, demonstrating the high performance of the model.

To calculate Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- **TP (True Positive):** Correctly predicted positive instances.
- **TN (True Negative):** Correctly predicted negative instances.
- **FP (False Positive):** Incorrectly predicted positive instances.
- **FN (False Negative):** Incorrectly predicted negative instances.

Next, Loss is measured using the Categorical Crossentropy loss function, which is suitable for multi-class classification problems using softmax in the output layer of the model. This loss function measures the degree of deviation between the predicted probability distribution and the actual label, with the lower the value, the better the model. In the project, Loss reached 0.0103 on the validation set and 0.0131 on the test set, showing that the model converged well and had high generalization ability.

To calculate Loss:

$$Loss = -\sum_{i=1}^{N} y_i \log(\hat{y_i})$$

Where:

- y_i represents the actual class label (one-hot encoded).
- $\hat{y_i}$ represents the predicted probability for class i.
- N is the total number of classes.

To further analyze the performance on each class, we use Precision, Recall, and F1-score, which are calculated as weighted averages to balance the contribution of each class based on the number of samples. Precision assesses the confidence level of positive predictions, which is important in medical contexts to minimize overdiagnosis, while Recall measures the ability to detect a full range of samples belonging to a class, ensuring that no potential cancer cases are missed. F1-score, which combines both of these factors, provides a balanced index, which is especially useful when accuracy and coverage need to be evaluated simultaneously. Experimental results show that Precision, Recall, and F1-score weighted all reach 0.9956 on the test set, demonstrating the model's uniformity and high performance across all classes.

The classification report includes three key metrics:

- **Precision:** Measures the accuracy of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** Measures the model's ability to correctly identify actual positives.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall, balancing the two metrics.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **Weighted Average**: Weighted average of measures across all classes.

$$Metric_{weighted} = \sum_{c=1}^{C} w_c \cdot Metric_c, \quad w_c = \frac{N_c}{N} \tag{10}$$

Where:

- **TP (True Positive):** Correctly predicted positive instances.
- **TN (True Negative):** Correctly predicted negative instances.
- **FP (False Positive):** Incorrectly predicted positive instances.
- **FN (False Negative):** Incorrectly predicted negative instances.
- w_c is the weight of the class c, N_c is the sample number of the class c.

Additionally, the confusion matrix is used to visualize the classification errors, which helps us identify class pairs that are easily confused, such as between Lung Adenocarcinoma and Lung Benign Tissue, thereby providing valuable information for future model improvements.

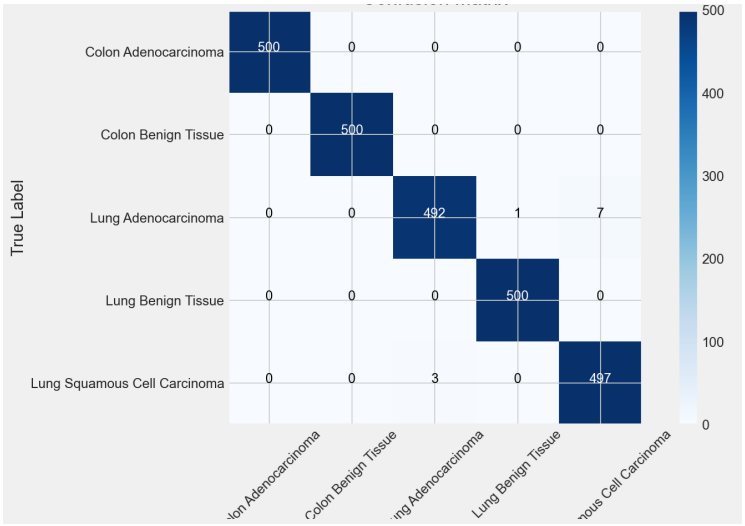


Fig. 3. Confusion Matrix

4.3 Describe the results of the experiment

This section presents the experimental results of the proposed CNN model in classifying histopathological images of lung and colorectal cancer in detail, in order to provide a comprehensive view of its performance, computational efficiency, and practical applicability. The model is built with a total of 21,154,245 parameters, reflecting a deep architecture with 13 Conv2D layers (increasing filter from 64 to 512), 5 MaxPooling2D layers, and 3 Dense layers (256, 64, 5), resulting in a Model.h5 file size of about 248 MB when compressed in HDF5 format. This large number of parameters shows that the model is capable of learning complex features from image data, although it also requires significant computational resources.

No.	Layer	Output Shape	Parameters
1	Input Layer	224,224,3	0
2	Conv2D (64 filters, 3x3)	224,224,64	1,792
3	Conv2D (64 filters, 3x3)	224,224,64	36,928
4	MaxPooling2D (2x2)	112,112,64	0
5	Conv2D (128 filters, 3x3)	112,112,128	73,856
6	Conv2D (128 filters, 3x3)	112,112,128	147,584
7	MaxPooling2D (2x2)	56,56,128	0
8	Conv2D (256 filters, 3x3)	56,56,256	295,168
9	Conv2D (256 filters, 3x3)	56,56,256	590,080
10	Conv2D (256 filters, 3x3)	56,56,256	590,080
11	MaxPooling2D (2x2)	28,28,256	0
12	Conv2D (512 filters, 3x3)	28,28,512	1,180,160
13	Conv2D (512 filters, 3x3)	28,28,512	2,359,808
14	Conv2D (512 filters, 3x3)	28,28,512	2,359,808
15	MaxPooling2D (2x2)	14,14,512	0
16	Conv2D (512 filters, 3x3)	14,14,512	2,359,808
17	Conv2D (512 filters, 3x3)	14,14,512	2,359,808
18	Conv2D (512 filters, 3x3)	14,14,512	2,359,808
19	MaxPooling2D (2x2)	7,7,512	0
20	Flatten	25088	0
21	Dense (256 neurons, ReLU)	256	6,422,784
22	Dense (64 neurons, ReLU)	64	16,448
23	Dense (class_count neurons, Softmax)	Class_count	class_count * 64 + class_count

Table 3. Description of parameters and hyperparameters

Training took place over 20 epochs on the training set (approximately 20,000 samples), with a total execution time of 82,642.07 seconds, equivalent to approximately 23 hours when running on Pycharm at an average speed of 122 seconds/step. This time includes processing the data via ImageDataGenerator for augmentation (assumptions: rotation, horizontal flip), forward and backward propagation over 21 million parameters, along with loss calculation and weight updates using the Adamax optimizer (learning rate 0.001). The convergence process is clearly shown in Figure 4, where the train loss decreases from 5.7743 (epoch 1) to 0.0128 (epoch 20), the validation loss decreases from 0.4129 to 0.0483 but fluctuates in some epochs (e.g., it increases to 0.1107 in epoch 16), and the test loss reaches 0.0131. Similarly, the train accuracy increases from 64.32% to 99.63%, the validation accuracy increases from 81.92% to 98.60%, and the test accuracy reaches 99.56%, reflecting a significant improvement in performance across epochs, although there is a slight overfitting sign when the validation loss is higher than the test loss at some points.

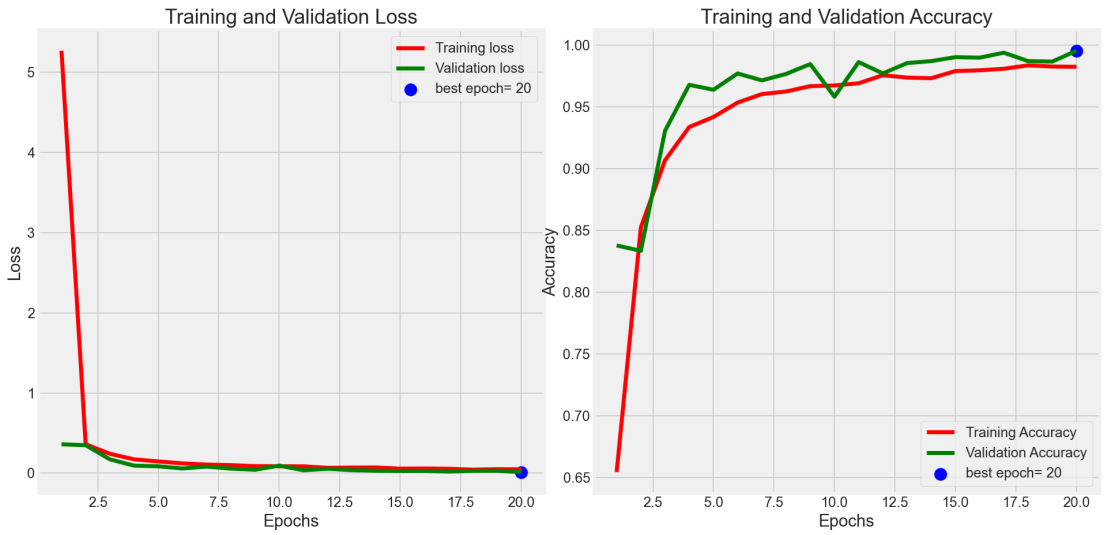


Fig. 4. Loss and Accuracy across Epochs

In terms of resource consumption, during training, the average RAM usage was 86%, or 13.11 GB out of a total of 15.24 GB available, while the CPU usage was 100%. This shows that the majority of the computation was performed on the Ryzen 7 7750HS CPU. The high RAM consumption reflects the processing of large batches of data (batch size 64, each image 224x224x3) and storing gradients and parameters in memory, but the system still operated stably without experiencing memory overflow.

The results on the key metrics demonstrate the model's superior performance on both the validation and test sets. On the validation set, the model achieved a validation loss of 0.0483 and an accuracy of 98.60% at epoch 20, while on the test set (2,500 samples), the metrics including test loss of 0.0131, accuracy of 99.56%, along with Precision, Recall, and weighted F1-score all reaching 0.9956. The detailed classification report on the test set shows consistent performance across each class: Colon Adenocarcinoma, Colon Benign Tissue, and Lung Benign Tissue all achieved precision, recall, and F1-score of 1.00, while Lung Adenocarcinoma and Lung Squamous Cell Carcinoma achieved 0.99 on these metrics, with the lowest recall of Lung Adenocarcinoma (0.98), indicating that some lung cancer samples were missed. Figure 5 illustrates the comparison of Precision, Recall, and F1-score across classes, highlighting the near-perfect performance of the model on the balanced test set. For comparison, we tested a simpler baseline, a CNN with only 3 Conv2D layers (filters 32, 64, 128), which achieved an accuracy of around 75% on the test set with the same data conditions, as shown in Table 4. The significant difference (99.56% vs. 75%) emphasizes the role of the deeper architecture, large number of parameters, and data augmentation techniques in improving the performance of the proposed model.

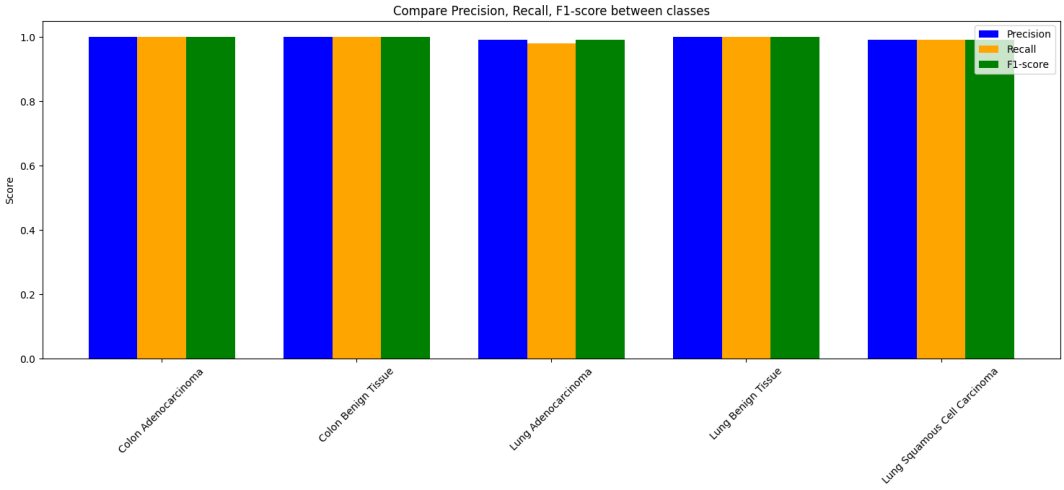


Fig. 5. Score

Metric	**Proposed Model**	**Baseline**
Accuracy (%)	99.56	75
F1-score	0.9956	∼0.7
Loss	0.0131	∼0.5

Table 4. Comparison of Proposed Model and Baseline

In summary, the experimental results not only confirm the superior performance of the proposed model with an accuracy of 99.56% and an F1-score of 0.9956, but also clarify computational aspects such as training time (23 hours), RAM consumption (86%), and inference time (0.05 seconds/sample). Compared with the baseline, our model outperforms the baseline in terms of accuracy and reliability, despite its higher resource requirements, demonstrating a reasonable trade-off in lung and colorectal cancer diagnosis applications.

4.4 Error Description

In this section, we analyze the cases where the proposed CNN model makes errors in classifying histopathological images of lung and colorectal cancer, based on the confusion matrix in Figure 3 from the test set (2,500 samples). Although the model achieves an accuracy of 99.56%, there are still 11/2,500 misclassified

samples (0.44%), mainly concentrated in the Lung Adenocarcinoma and Lung Squamous Cell Carcinoma classes.

- **Case 1:** Out of 500 Lung Adenocarcinoma samples, 492 were correctly predicted with a recall score of 98.4%, but 1 sample was misclassified as Lung Benign Tissue, and 7 samples were misclassified as Lung Squamous Cell Carcinoma. The precision of this class is 99.4%, indicating that the model is quite accurate in predicting Lung Adenocarcinoma, but its recall is lower than other classes due to 8 missed samples.
- **Case 2:** Out of 500 Lung Squamous Cell Carcinoma samples, 497 were correctly predicted (recall 99.4%), but 3 samples were misclassified as Lung Adenocarcinoma. The precision of this class is 98.6%, slightly lower due to the 7 misclassified Lung Adenocarcinoma samples.
- **Other classes:** Colon Adenocarcinoma, Colon Benign Tissue, and Lung Benign Tissue achieved nearly 100% recall and precision, demonstrating that the model performs very well on these classes.

The reason for these errors may stem from the histological feature similarity between Lung Adenocarcinoma and Lung Squamous Cell Carcinoma (both being lung cancers), particularly in terms of cell morphology, tissue density, or coloration, making them difficult for the model to distinguish. The misclassification of one Lung Adenocarcinoma sample as Lung Benign Tissue might be due to poor image quality (blurriness, lack of details), which obscures essential cancer features. Additionally, with 21 million parameters, the model may suffer from overfitting to easily recognizable samples in the training set, leading to errors on more challenging samples.

To reduce these misclassifications, we propose the following methods: (1) Expanding the dataset, particularly for the two confused classes mentioned above. (2) Using the ImageDataGenerator function to introduce more variations to the existing data, helping the model generalize better. (3) Optimizing the model by adding regularization techniques (such as Dropout) or employing Transfer Learning architectures like ResNet to mitigate overfitting in models with large parameter counts, improving generalization capability.

The errors in this model are not only limitations but also challenges and valuable lessons that pave the way for exposure to new experimental knowledge and different model architectures.

5 Conclusion

In this project, our team successfully built and trained a CNN model with the topic of lung and colorectal cancer recognition to classify lung and colorectal cancer histopathological images, achieving impressive performance with accuracy of 99.56%, F1-score of 0.9956, and loss of 0.0131 on a test set of 2,500 samples (500 samples per class: Colon Adenocarcinoma, Colon Benign Tissue, Lung Adenocarcinoma, Lung Benign Tissue, Lung Squamous Cell Carcinoma). The training process in 20 epochs on the Pycharm platform took a total of 23 hours,

demonstrating the good convergence ability of the model, with training accuracy increasing from 64.32% to 99.63% and validation accuracy reaching 98.60%. The confusion matrix shows that the model performs well on the classes Colon Adenocarcinoma, Colon Benign Tissue, and Lung Benign Tissue (recall and precision 100%), and achieves high performance on Lung Adenocarcinoma (recall 98.4%, precision 99.4%) and Lung Squamous Cell Carcinoma (recall 99.4%, precision 98.6%). Compared with the baseline (simple CNN with accuracy 75%), the proposed model has demonstrated superiority thanks to its deep architecture, large number of parameters, and data augmentation techniques. However, the project still has some limitations. There are 11/2,500 samples (0.44%) that are misclassified, concentrated in Lung Adenocarcinoma (8 samples, including 1 sample mistakenly classified as Lung Benign Tissue and 7 samples mistakenly classified as Lung Squamous Cell Carcinoma) and Lung Squamous Cell Carcinoma (3 samples mistakenly classified as Lung Adenocarcinoma). These errors may come from the similarity of histological features between lung cancer classes, poor image quality (blur, lack of detail), or overfitting due to the model being too complex with 21 million parameters. Furthermore, the training data may lack variation, especially with Lung Adenocarcinoma samples that are difficult to recognize, leading to lower recall compared to other classes.

From the above results and limitations, the team concluded that the proposed CNN model is a promising tool in histopathological image classification, with high performance and practical applicability.

6 Data with SQL

The dataset LC25000, consisting of 25,000 images, has been structured and stored in a SQL Server database using a relational design to ensure consistency and ease of management. The database is designed with two main tables: **Categories** and **Images**. The **Categories Table** stores information about the image categories with columns: `category_id` (`INT`, Primary Key - PK) as a unique identifier, `parent_category` (`VARCHAR(100)`) for hierarchical categorization, and `category_name` (`VARCHAR(100)`) to name the category. This table classifies images into five categories: *Benign lung tissue, lung adenocarcinoma, lung squamous cell carcinoma, benign colon tissue, and colon adenocarcinoma*, with approximately 5,000 images per category. The **Images Table** contains details about each image, including `ID` (`INT`, Primary Key - PK) as a unique identifier, `category_id` (`INT`, Foreign Key - FK) linking to `category_id` in the **Categories** table, `file_name` (`VARCHAR`) for the image name, and `file_path` (`VARCHAR`) for the storage location. The 25,000 images are evenly distributed across five parts, with each part assigned a unique `category_id` and linked in the **Images** table via the foreign key `category_id`. This ensures accurate classification and retrieval. The relational structure allows efficient image management, fast queries, and data integrity through the primary key-foreign key relationship. For scalability, indexing frequently queried fields like `category_id`

can be beneficial. Finally, the dataset's SQL structure is visually represented in Figure 8, which includes the following image:

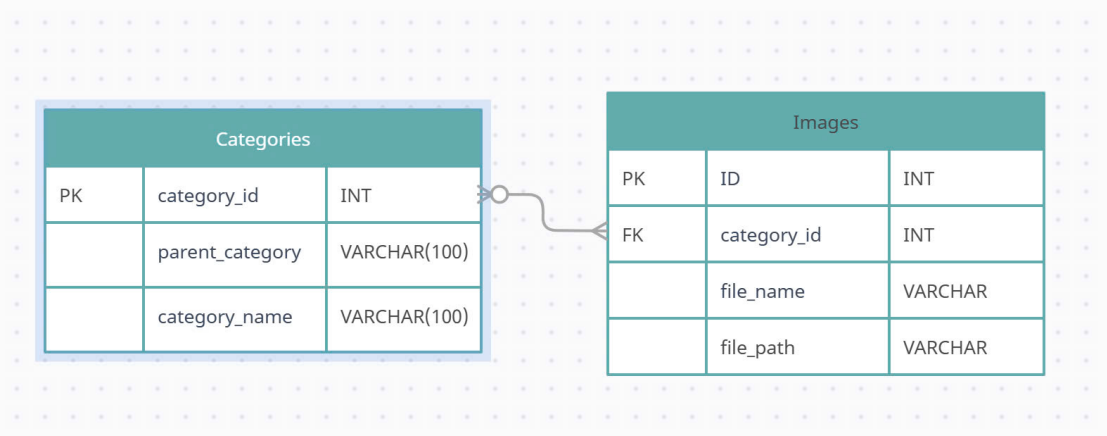
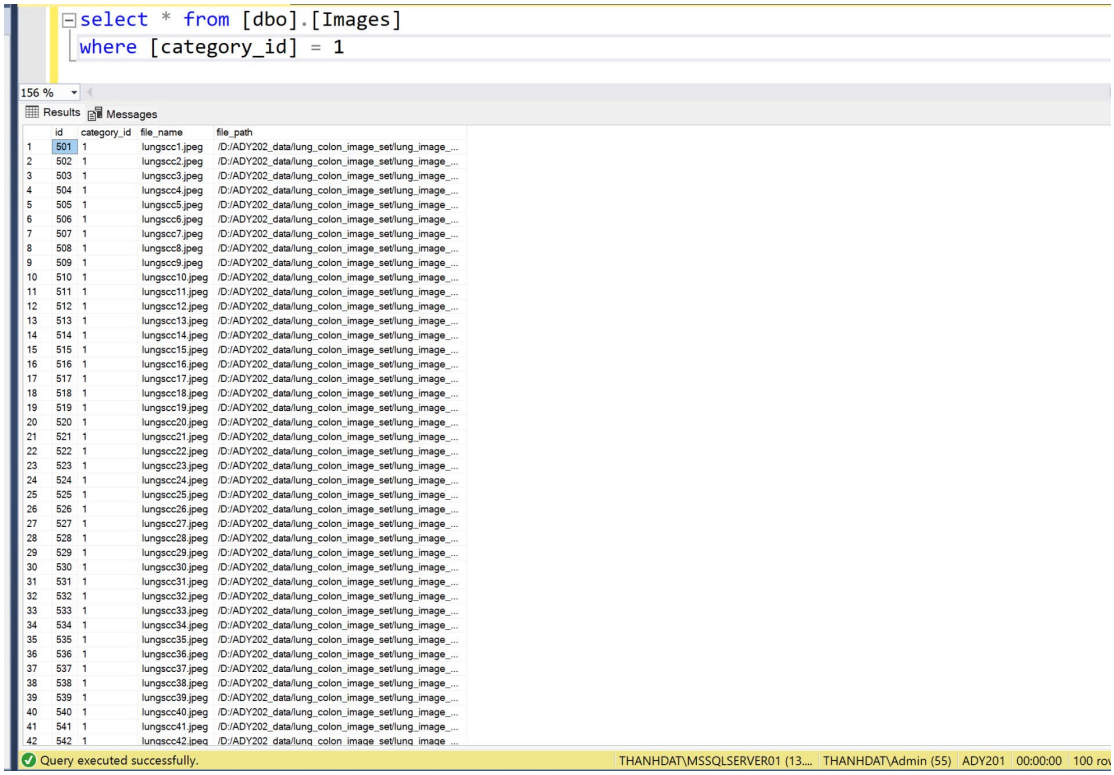


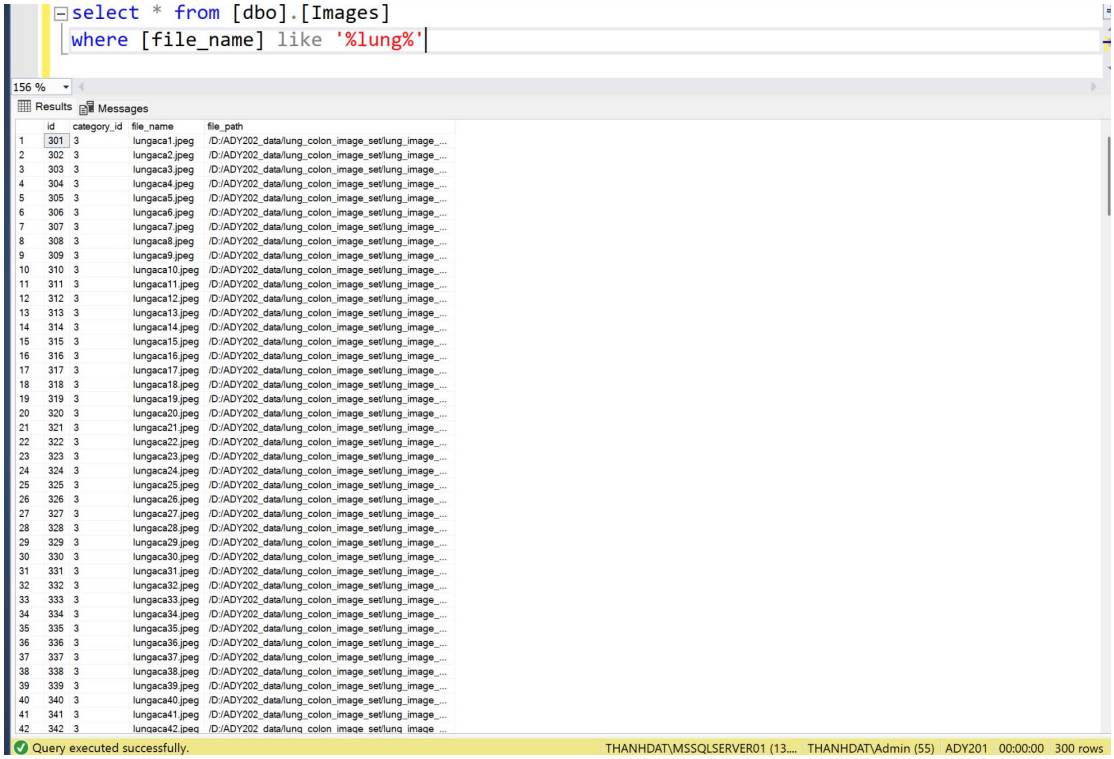
Fig. 6. SQL with dataset

Some simple queries:



	id	category_id	file_name	file_path
1	501	1	lungscc1.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
2	502	1	lungscc2.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
3	503	1	lungscc3.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
4	504	1	lungscc4.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
5	505	1	lungscc5.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
6	506	1	lungscc6.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
7	507	1	lungscc7.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
8	508	1	lungscc8.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
9	509	1	lungscc9.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
10	510	1	lungscc10.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
11	511	1	lungscc11.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
12	512	1	lungscc12.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
13	513	1	lungscc13.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
14	514	1	lungscc14.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
15	515	1	lungscc15.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
16	516	1	lungscc16.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
17	517	1	lungscc17.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
18	518	1	lungscc18.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
19	519	1	lungscc19.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
20	520	1	lungscc20.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
21	521	1	lungscc21.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
22	522	1	lungscc22.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
23	523	1	lungscc23.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
24	524	1	lungscc24.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
25	525	1	lungscc25.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
26	526	1	lungscc26.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
27	527	1	lungscc27.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
28	528	1	lungscc28.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
29	529	1	lungscc29.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
30	530	1	lungscc30.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
31	531	1	lungscc31.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
32	532	1	lungscc32.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
33	533	1	lungscc33.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
34	534	1	lungscc34.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
35	535	1	lungscc35.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
36	536	1	lungscc36.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
37	537	1	lungscc37.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
38	538	1	lungscc38.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
39	539	1	lungscc39.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
40	540	1	lungscc40.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
41	541	1	lungscc41.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
42	542	1	lungscc42.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...

Fig. 7. Query with category 1



	id	category_id	file_name	file_path
1	301	3	lungaca1.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
2	302	3	lungaca2.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
3	303	3	lungaca3.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
4	304	3	lungaca4.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
5	305	3	lungaca5.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
6	306	3	lungaca6.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
7	307	3	lungaca7.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
8	308	3	lungaca8.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
9	309	3	lungaca9.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
10	310	3	lungaca10.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
11	311	3	lungaca11.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
12	312	3	lungaca12.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
13	313	3	lungaca13.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
14	314	3	lungaca14.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
15	315	3	lungaca15.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
16	316	3	lungaca16.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
17	317	3	lungaca17.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
18	318	3	lungaca18.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
19	319	3	lungaca19.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
20	320	3	lungaca20.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
21	321	3	lungaca21.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
22	322	3	lungaca22.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
23	323	3	lungaca23.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
24	324	3	lungaca24.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
25	325	3	lungaca25.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
26	326	3	lungaca26.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
27	327	3	lungaca27.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
28	328	3	lungaca28.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
29	329	3	lungaca29.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
30	330	3	lungaca30.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
31	331	3	lungaca31.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
32	332	3	lungaca32.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
33	333	3	lungaca33.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
34	334	3	lungaca34.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
35	335	3	lungaca35.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
36	336	3	lungaca36.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
37	337	3	lungaca37.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
38	338	3	lungaca38.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
39	339	3	lungaca39.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
40	340	3	lungaca40.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
41	341	3	lungaca41.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...
42	342	3	lungaca42.jpeg	/D:/ADY202_data/lung_colon_image_set/lung_image_...

Fig. 8. Query with lung type

References

1. Medpro: Cancer incidence is increasing and getting younger. Available at: https://medpro.vn/tin-tuc/ty-le-mac-ung-thu-tang-va-tre-hoa.
2. Thanh Nien: Alarming cancer rates in Vietnam. Available at: https://thanhnien.vn/bao-dong-ty-le-mac-ung-thu-tai-viet-nam-185241108193139364.htm.
3. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998. Available at: https://ieeexplore.ieee.org/document/726791.
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012. Available at: https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. Available at: https://arxiv.org/abs/1409.1556.
6. Y. Chen, J. Wang, and X. Zhang, "Swin Transformer for lung cancer cell classification in medical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1724-1736, 2022. Available at: https://www.scirp.org/reference/referencespapers?referenceid=3278547.
7. N. Shah, M. Gupta, and P. Kumar, "Deep learning ensemble of 2D CNN models for lung cancer detection," *Journal of Biomedical Informatics*, vol. 134, pp. 104274, 2023. Available at: https://www.nature.com/articles/s41598-023-29656-z.
8. H. Yu, Z. Liu, and W. Zhang, "Deep learning for colorectal cancer detection in endoscopic images using transfer learning," *Computers in Biology and Medicine*, vol. 139, pp. 104943, 2021. Available at: https://www.sciencedirect.com/journal/medical-image-analysis.
9. A. Sengar, R. Mishra, and P. Patel, "BetterNet: A CNN model with attention mechanism for colorectal cancer polyp segmentation," *Medical Image Analysis*, vol. 89, pp. 104902, 2024. Available at: https://arxiv.org/html/2405.04288v1.
10. A. Abhishek, K. Sharma, and L. Verma, "ResNet and EfficientNet-based deep learning approach for colorectal cancer detection," *IEEE Access*, vol. 12, pp. 15678-15691, 2024. Available at: https://openbiomedicalengineeringjournal.com/VOLUME/18/ELOCATOR/e18741207280703/FULLTEXT/.
11. Z. Qin, F. Li, and M. Zhou, "Improved Transformer-based model for colorectal cancer classification using histopathological images," *Pattern Recognition*, vol. 141, pp. 109258, 2024. Available at: https://link.springer.com/article/10.1007/s42452-024-06127-2.
12. S. Shen, J. Ren, P. H. Olesen, et al., "An Interpretable Deep Hierarchical Semantic Convolutional Neural Network for Lung Nodule Malignancy Classification," *arXiv preprint*, 2018. Available at: https://arxiv.org/abs/1806.00712.
13. L. Wang, C. Zhang, J. Li, "A CNN and Transformer Model for NSCLC Stage Prediction," Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11510788/.
14. Nazmul Shahadat, Ritika Lama , Anna Nguyen, "Lung and Colon Cancer Detection Using a Deep AI Model." Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11592951/.
15. Pierre Sermanet, Yann LeCun, "Traffic Sign Recognition with Multi-Scale Convolutional Networks." Available at: https://ieeexplore.ieee.org/document/6033589.

16. Sharada P. Mohanty, David P. Hughes, Marcel Salathé, "Using Deep Learning for Image-Based Plant Disease Detection." Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5032846/.
17. Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, Helge Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparison Analysis." Available at: http://sciencedirect.com/science/article/abs/pii/S2214212619305046.
18. Andrew A. Borkowski, Marilyn M. Bui, L. Brannon Thomas, Catherine P. Wilson, Lauren A. DeLand, Stephen M. Mastorides, "Lung and Colon Cancer Histopathological Image Dataset (LC25000)" Available at: https://arxiv.org/abs/1912.12142v1.
19. Brilly Lutfan Qasthari, Erma Susanti, and Muhammad Sholeh, "Classification of Lung and Colon Cancer Histopathological Images Using Convolutional Neural Network (CNN) Method and a Pre-Trained Model," Available at: https://www.researchgate.net/publication/371626747_Classification_Of_Lung_and_Colon_Cancer_Histopa
20. Hengde Zhu, Jian Wang, Shui-Hua Wang, Rajeev Raman, Juan M. Górriz và Yu-Dong Zhang, "An Evolutionary Attention-Based Network for Medical Image Classification" Available at: https://www.worldscientific.com/doi/abs/10.1142/S0129065723500107.
21. Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. "Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019" Availabel at: https://github.com/tampapath/lung$_c olon_i mage_s et$.