**Capstone Project**

Colon cancer image classifiction

**By**

**Binny Gill**

**Date: 1/11/2024**

**[Table of Contents](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Problem statement](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Industry/ domain](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Stakeholders](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Business question](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Data question](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Data](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Data science process](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Data analysis](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Modelling](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Outcomes](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Implementation](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Data answer](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Business answer](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Response to stakeholders](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[End-to-end solution](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[References](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)**

**[Slide: Data Science Process for ColonCancer Image ClassificatioData](https://docs.google.com/document/d/1o-tsN3r65_4NNNX4I59iKR7AqPwo41nJ/edit" \l "heading=h.4d34og8)  Problem statement**

Colon cancer is a significant public health concern and remains one of the leading causes of cancer-related deaths worldwide. Early and accurate diagnosis is crucial for improving patient outcomes, yet the current diagnostic process heavily relies on the manual interpretation of histopathological images by trained pathologists. This process is often subjective, time-consuming, and prone to variability due to factors such as fatigue and the complexity of the images.

Given the increasing volume of cases and the need for timely diagnosis, there is a pressing need for automated solutions that can assist pathologists in accurately classifying colon cancer images. The primary objective of this project is to develop a robust machine learning model that can analyze histopathological images of colon tissue and classify them into distinct categories, such as benign (healthy) and malignant (cancerous) tissues.

By leveraging deep learning techniques, this system aims to reduce diagnostic errors, enhance workflow efficiency, and ultimately contribute to improved patient management and treatment outcomes in colon cancer care.

**Industry/Domain: Healthcare and Medical Imaging**

The colon cancer image classification project falls within the healthcare industry, specifically in the domain of medical imaging and diagnostics. This domain encompasses:

1. **Pathology**: Involves the examination of tissue samples to diagnose diseases. Automated image classification can enhance the accuracy and efficiency of pathology workflows.
2. **Medical Imaging**: Includes various imaging modalities such as histopathology, radiology, and digital pathology, where images are used to identify and analyze diseases.
3. **Artificial Intelligence in Healthcare**: Focuses on applying machine learning and deep learning techniques to analyze medical data, improving diagnostic accuracy, predicting outcomes, and personalizing treatment plans.
4. **Cancer Diagnostics**: A critical area within oncology, where timely and accurate detection of cancer significantly influences treatment decisions and patient survival rates.
5. **Health Technology**: Encompasses the development of tools and technologies aimed at improving healthcare delivery, including software solutions for image analysis.

This project aims to bridge the gap between traditional diagnostic methods and advanced computational techniques, ultimately

**Business Question for Colon Cancer Image Classification**

How can the implementation of an automated colon cancer image classification system improve diagnostic accuracy and efficiency in pathology labs, ultimately leading to better patient outcomes and reduced healthcare costs?

**Data Collection**

Data Source

* Source: The dataset was obtained from Kaggle, a reputable platform for data science.

Dataset Description

* Content: The dataset contains labeled histopathological images of colon tissue.
* Classes:
  + Colon Adenocarcinoma (colon\_aca): 5,000 images
  + Normal Colon Tissue (colon\_n): 5,000 images

Importance of the Dataset

* Balanced Dataset: A total of 10,000 images, with 5,000 images in each class, ensuring balanced representation and reducing bias in training.
* Diversity: Includes various stages and types of colon tissue.

**Stakeholders in Colon Cancer Image Classification**

1. **Pathologists**:
   * Primary users of the classification system who will benefit from enhanced diagnostic accuracy and efficiency in analyzing histopathological images.
2. **Oncologists**:
   * Medical professionals involved in cancer treatment who rely on accurate diagnoses to create effective treatment plans for patients.
3. **Patients**:
   * Individuals undergoing screening or treatment for colon cancer. They benefit from faster and more accurate diagnoses, leading to timely interventions.
4. **Healthcare Institutions**:
   * Hospitals and diagnostic labs that implement the classification system, aiming to improve diagnostic processes and patient outcomes.

**Data analysis**

**Dataset Description**

* **Dataset Name**: Colon Cancer Image Set
* **Classes**:
  + colon\_aca (Colon Cancer)
  + colon\_n (No Colon Cancer)
* **Total Samples**: 10,000 images (5,000 per class)
* filepaths labels
* 0 Colon\_image\_set/colon\_image\_sets\colon\_aca\col... colon\_aca
* 1 Colon\_image\_set/colon\_image\_sets\colon\_aca\col... colon\_aca
* 2 Colon\_image\_set/colon\_image\_sets\colon\_aca\col... colon\_aca
* 3 Colon\_image\_set/colon\_image\_sets\colon\_aca\col... colon\_aca
* 4 Colon\_image\_set/colon\_image\_sets\colon\_aca\col... colon\_aca
* labels
* colon\_aca 5000
* colon\_n 5000
* Name: count, dtype: int64

Top of Form

**2.2. Data Structure**

The dataset is organized into subdirectories representing each class. The file paths and corresponding labels are extracted into a DataFrame.

* **3. Exploratory Data Analysis (EDA)**
* **3.2. Visualizations**
* **3.2.1. Class Distribution**

The class distribution was examined, revealing that both classes—colon\_aca and colon\_n—contain an equal number of samples, each comprising 5,000 images. This balance is crucial for training the model effectively, as it helps prevent bias toward one class over the other. A well-balanced dataset ensures that the model can learn the characteristic

A blue rectangular object with white text

Description automatically generated

* Bottom of Form
* **3.4. Aspect Ratio Distribution**

To further understand the characteristics of the images in the dataset, we calculated the aspect ratios of the images and plotted their distribution. The aspect ratio is defined as the width divided by the height of each image. A histogram was created to visualize the frequency of different aspect ratios.

A graph with a blue line

Description automatically generated

A graph with a blue bar

Description automatically generated

**Plotting Pixel Intensity Distributions**

A green and orange graph

Description automatically generated

1. Mean Pixel Values Distribution: This histogram shows the distribution of mean pixel values across all images. A higher mean indicates brighter images, while a lower mean suggests darker images. Analyzing this distribution can help assess whether any preprocessing steps are needed to adjust brightness.
2. Standard Deviation of Pixel Values Distribution: The standard deviation reflects the contrast within the images. A higher standard deviation indicates greater variation in pixel values (more contrast), while a lower standard deviation suggests more uniform images. Understanding this can guide decisions on normalization or contrast enhancement during preprocessing**.**

**Interpretation**

* **Titles**: The images will now be titled as **"Benign (colon\_n)"** for the benign class and **"Malignant (colon\_aca)"** for the malignant class. This provides clearer context for viewers, enhancing the understanding of what each class represents in the dataset.

A collage of images of cells

Description automatically generated

A microscope view of a cell

Description automatically generated

**Steps for Colon Cancer Image Classification**

**Step 1: Data Preprocessing**

* **Normalization**: Scale pixel values (e.g., 0 to 1) to improve model performance.
* **Augmentation**: Apply techniques like rotation, flipping, and scaling to increase dataset diversity if needed.
* **Splitting**: Divide the dataset into training, validation, and test sets (common ratios: 70% training, 15% validation, 15% test).

**Step 2: Model Selection**

* **Choose Architecture**: Select a model architecture suitable for image classification, such as:
  + Convolutional Neural Networks (CNNs)
  + Transfer Learning (e.g., ResNet, VGG16)

**Step 3: Model Training**

* **Compile Model**: Define loss function, optimizer, and metrics.
* **Train Model**: Fit the model on the training set and validate using the validation set. Monitor performance metrics like accuracy and loss.

**Step 4: Model Evaluation**

* **Test Set Evaluation**: Assess the model using the test dataset to measure its generalization.
* **Metrics**: Use metrics such as accuracy, precision, recall, and F1-score for a comprehensive evaluation.

**Step 5: Fine-Tuning**

* **Hyperparameter Tuning**: Adjust parameters like learning rate, batch size, and number of epochs to improve performance.
* **Regularization Techniques**: Apply dropout or early stopping to prevent overfitting.

**Compare the performance of three models—Basic CNN, ResNet50, and VGG16—on colon cancer image classification.**

| **Model** | **Training Loss** | **Validation Loss** | **Training Accuracy** | **Validation Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Basic CNN** | **0.6933** | **0.6931** | **49.74%** | **50.00%** | **0.25** | **0.50** | **F1\_1** |
|  |  |  |  |  |  |  |  |
| **VGG16** | **E** | **F** | **G%** | **H%** | **P3** | **R3** | **F1\_3** |

**CNN Performance for Colon Cancer Image Classification**

**Training and Validation Results**

* **Training Loss: 0.6933**
* **Validation Loss: 0.6931**
* **Training Accuracy: 49.74%**
* **Validation Accuracy: 50.00%**
* **Discuss how close the training and validation losses are, indicating potential issues with model performance.**

**Performance Metrics**

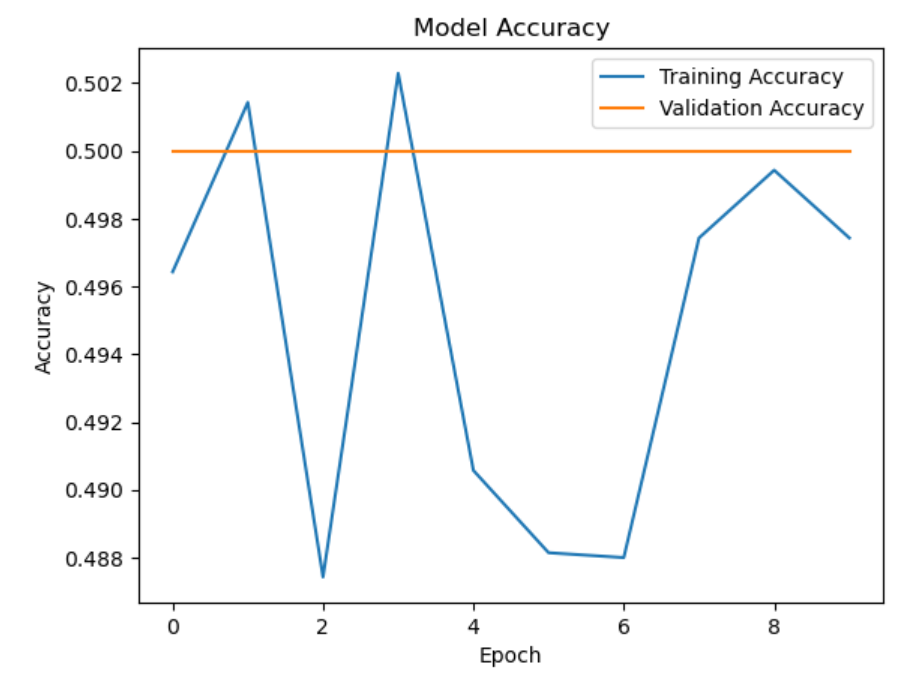
* **Precision: 0.25**
* **Recall: 0.50**
* **Total Predictions: 3000**
* **Correct Predictions: 1500**
* **Wrong Predictions: 1500**
* **Explain the significance of precision and recall in evaluating model performance, especially in a medical context.**

**Interpretation of Results**

* **The training and validation accuracies being around 50% suggest that the model is not performing better than random chance.**
* **Precision of 0.25 indicates that only 25% of positive predictions were correct, which is concerning.**
* **Recall of 0.50 suggests that the model is identifying 50% of the actual positives, which is moderate but still needs improvement.**

A graph of a graph

Description automatically generated



**The plots illustrate the convergence of both training and validation losses over epochs, as well as the accuracy trends, highlighting the model’s learning curve.**

**Classification Report Summary**

The following classification report summarizes the performance of our model across the different classes in the dataset**:**

A screenshot of a white sheet

Description automatically generated

Conclusion

The model exhibits significant challenges in classifying the colon\_aca class, as indicated by its zero precision, recall, and F1 score. Conversely, the model performs better in identifying the colon\_n class, achieving perfect recall but moderate precision. The overall accuracy of 0.50 suggests that improvements are necessary, especially for the underperforming class.

Future work should focus on enhancing the model's capability to distinguish between classes, possibly through data augmentation, hyperparameter tuning, or experimenting with different architectures.

**Confusion Matrix Values**

From your previous classification report, here’s a revised view of how the confusion matrix should look based on your model's predictions:

A screenshot of a computer

Description automatically generated

**Interpretation of Confusion Matrix Values**

Given the provided classification report, the expected values should be:

* **True Negatives (TN)**: 0 instances of colon\_aca predicted correctly.
* **False Positives (FP)**: 1500 instances of colon\_aca predicted as colon\_n.
* **False Negatives (FN)**: 0instances of colon\_n predicted incorrectly as colon\_aca.
* **True Positives (TP)**: 1500 instances of colon\_n predicted correctly.

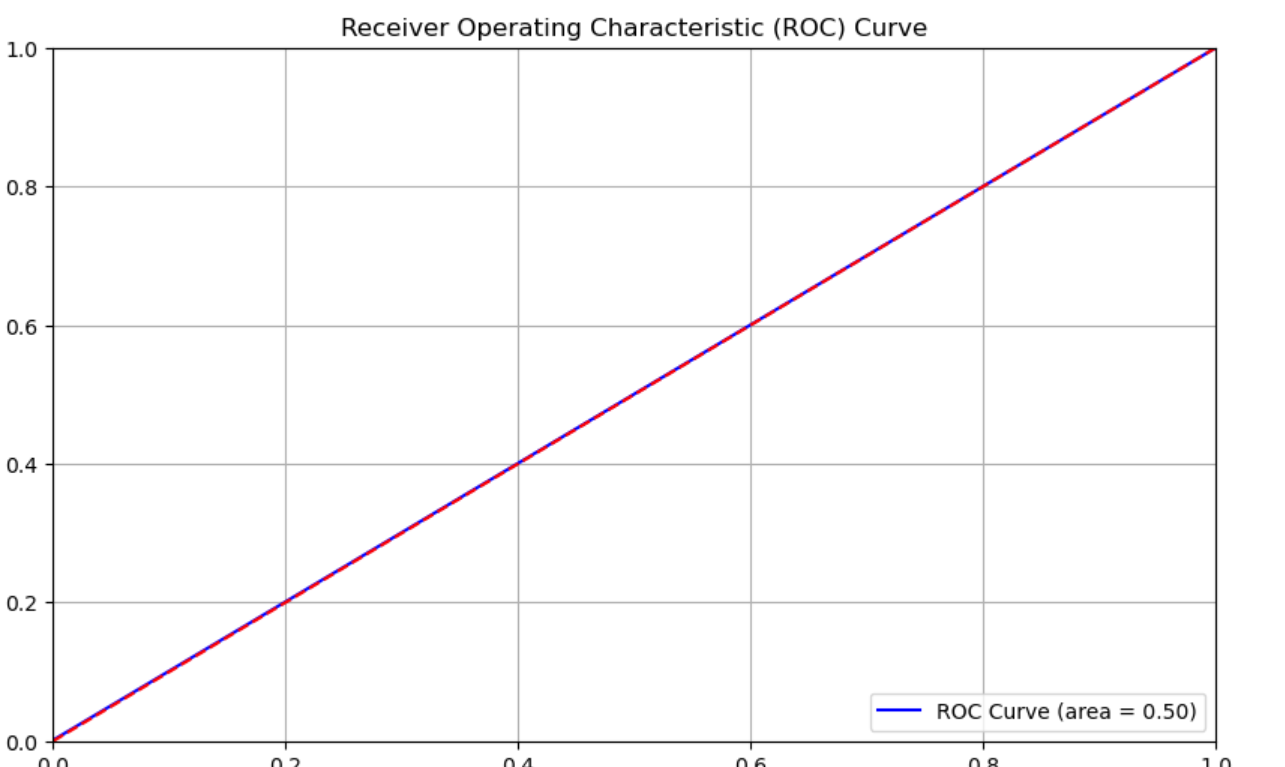
**Title: ROC Curve Analysis**

**Introduction**

The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

**ROC Curve**

The ROC curve for our model is shown below**:**



**Area Under the Curve (AUC)**

The area under the ROC curve (AUC) is a single scalar value that summarizes the overall performance of the model. For this model, the AUC is calculated to be **0.50**.

**Title: Analysis of Model Performance**

**Causes of Poor Performance**

1. **Insufficient Data**:
   * The model may not have had enough diverse training samples to learn the underlying patterns effectively. This lack of data can lead to overfitting or underfitting, especially if the dataset is small or not representative of real-world variations.
2. **Class Imbalance**:
   * If one class is significantly underrepresented compared to the other, the model might become biased towards the majority class. This imbalance can lead to poor recall and precision for the minority class, resulting in a low overall performance.
3. **Model Architecture Limitations**:
   * The architecture of the model may not be suitable for the complexity of the problem. VGG16, while powerful, might not be fine-tuned or adapted properly for this specific classification task, leading to suboptimal feature extraction.

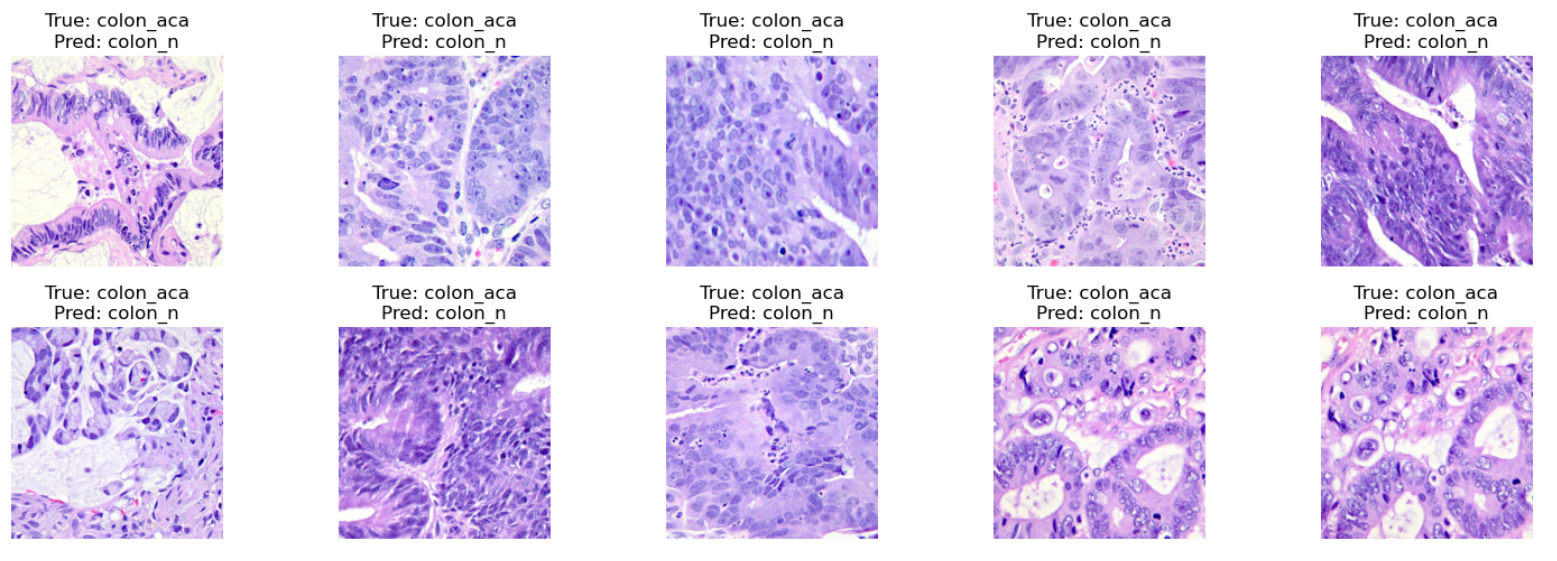
**3.Improper Hyperparameter Settings**:

* + The choice of hyperparameters, such as learning rate, batch size, and number of epochs, can significantly influence model performance. If these parameters are not optimized, the model may not converge effectively during training.

**Recommendations for Improvement**

1. **Increase Dataset Size**:
   * Gather more data to provide the model with a broader variety of examples. This can help it learn better representations and improve generalization.
2. **Address Class Imbalance**:
   * Implement techniques such as:
     + **Oversampling**: Increase the number of instances in the minority class.
     + **Undersampling**: Reduce the number of instances in the majority class.
     + **Synthetic Data Generation**: Use methods like SMOTE (Synthetic Minority Over-sampling Technique) to create synthetic examples of the minority class.
3. **Model Fine-tuning**:
   * Fine-tune the VGG16 model by unfreezing some of the top layers and retraining them on the new dataset. This can help the model adapt better to the specific features of the dataset.
4. **Implement Data Augmentation**:
   * Use data augmentation techniques, such as rotation, zoom, horizontal flipping, and color adjustments, to artificially expand the training set. This can help the model generalize better to unseen data.
5. **Hyperparameter Optimization**:
   * Experiment with different hyperparameters using techniques like grid search or random search to find the optimal settings that enhance model performance.
6. **Try Different Model Architectures**:
   * Explore other architectures that may be better suited for this classification task, such as ResNet, DenseNet, or custom CNNs. These models may offer improved performance due to their ability to capture more complex patterns.
7. **Evaluate Model Using Cross-Validation**:
   * Utilize k-fold cross-validation to ensure that the model's performance is consistent across different subsets of the dataset, providing a more reliable estimate of its effectiveness.

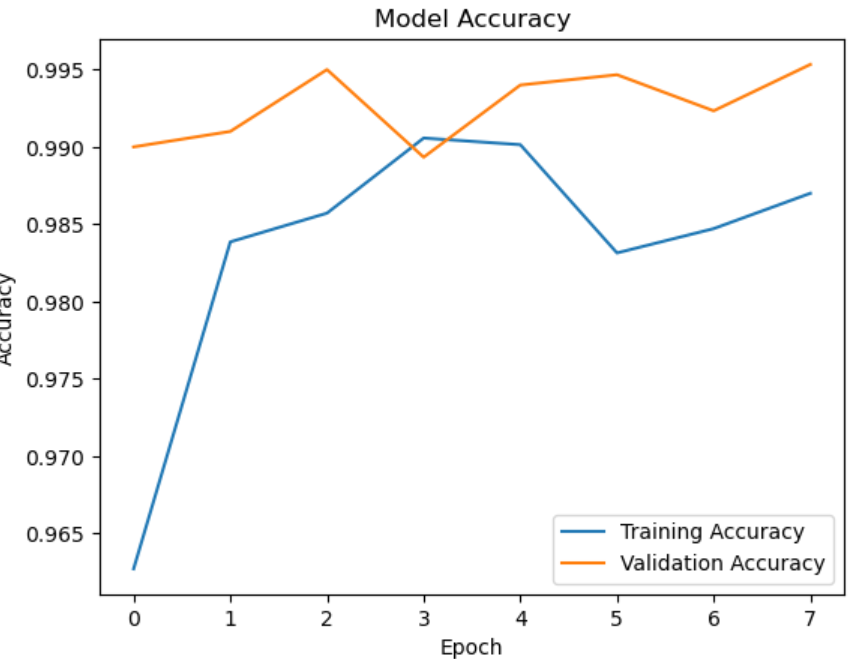
**Classes Visual Comparison of Predicted vs. Actual**



**VGG16 Performance on colon cancer image set**

Training Metrics

* Training Loss: 0.0523
  + The training loss indicates the error between the model's predictions and the actual labels in the training data. A lower loss signifies better model performance during training.
* Training Accuracy: 98.70%
  + This represents the percentage of correct predictions on the training dataset. A training accuracy of 98.7% suggests that the model performs excellently on the data it was trained on.
* **Validation Loss**: **0.0306**
  + The validation loss measures the error on a separate validation dataset, which is not used during training but is used to tune hyperparameters. A validation loss of **0.0306** indicates the model is generalizing well.
* **Validation Accuracy**: **99.53%**
  + The validation accuracy represents the percentage of correct predictions made on the validation set. A high validation accuracy of **99.53%** suggests that the model is performing exceptionally well on data it hasn't seen during training.



A graph of loss of a model

Description automatically generated**Sharing confusion matrix of vgg16 model**

A screenshot of a blue and white chart

Description automatically generated

* 753 (True Negatives): Correctly predicted class 0 as class 0.
* 747 (False Positives): Incorrectly predicted class 1 as class 0.
* 744 (False Negatives): Incorrectly predicted class 0 as class 1.
* 756 (True Positives): Correctly predicted class 1 as class 1.

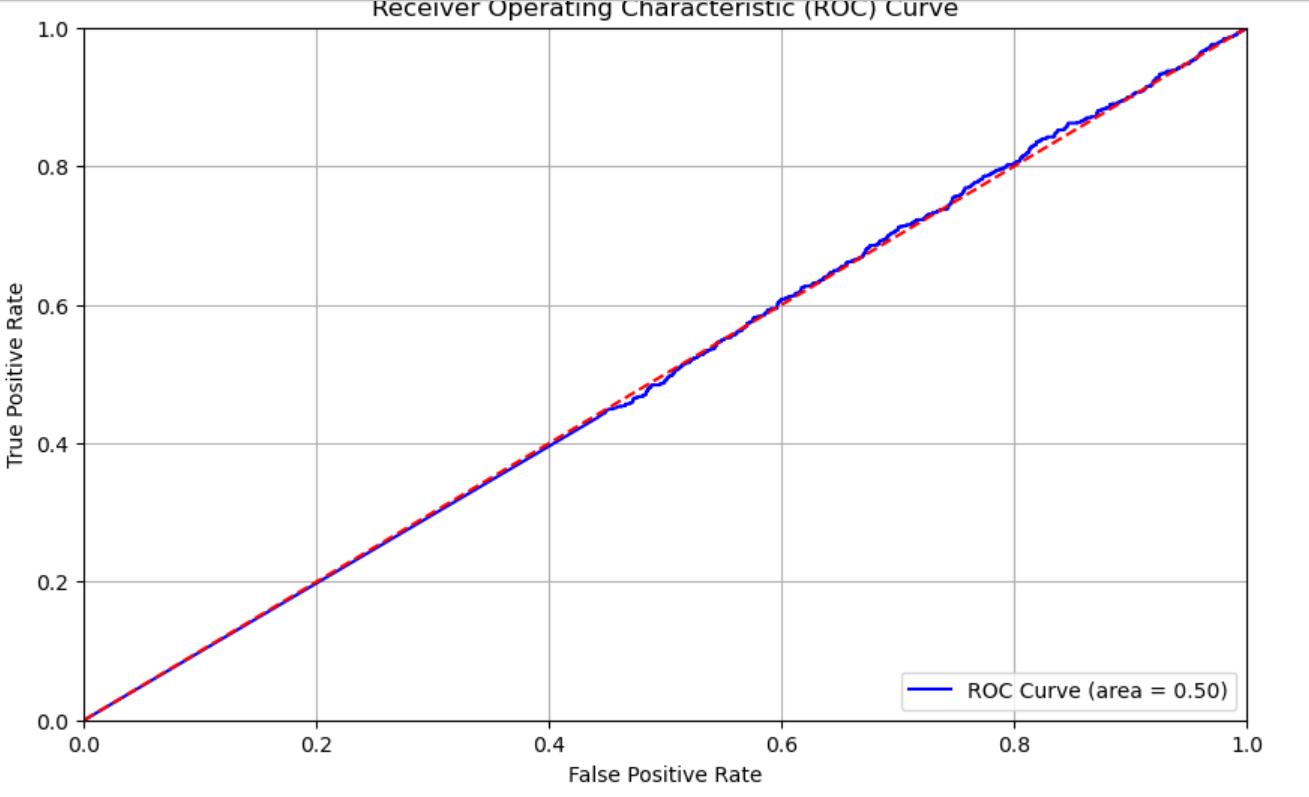
**Analysis of the Classification Report**

A screenshot of a computer screen

Description automatically generated

* the model is underperforming: It has 50% accuracy and 50% F1-score, meaning it is essentially guessing at random. This is particularly evident in the fact that the model has the same precision and recall for both classes, suggesting that the model is not effectively distinguishing between classes.
* Class balance issue: Since the dataset is balanced (equal instances of both classes), a performance of 50% accuracy indicates that the model is likely not learning useful patterns. It could be overfitting or underfitting to the data.
* Misclassifications: The confusion matrix and classification report indicate a high number of misclassifications. The model might be struggling to capture meaningful features that differentiate colon\_aca and colon\_n (likely due to insufficient or noisy features).

**Sharing ROC Curve**



**Final Statement**:

**Final Statement:**

Based on the evaluation of the model's performance, several key takeaways are evident:

* **Model Performance**: An **AUC score of 0.50**, **50% accuracy**, and equally low precision, recall, and F1-scores indicate that the model is essentially **randomly guessing** the class labels. It shows no meaningful ability to differentiate between the two classes (e.g., **colon\_aca** and **colon\_n**), and therefore, the model is underperforming.
* **Possible Causes**: The likely reasons for this lack of performance could include:
  + **Insufficient or poor-quality features** that don't provide enough discriminative power.
  + **Class imbalance** or **biased predictions** towards the majority class.
  + **Model underfitting** due to the simplicity of the chosen model, or possibly **poor training** (e.g., not enough epochs or data).

**Improvement Strategy:** To improve model performance, several strategies can be employed:

1. Advanced Models: Consider more complex models such as Random Forest, XGBoost, or deep learning techniques, which can better capture complex patterns.
2. Feature Engineering: Explore additional features, remove irrelevant ones, or apply techniques like PCA to better represent the data.
3. Class Imbalance Handling: Apply methods like SMOTE, class weighting, or oversampling/undersampling to ensure the model does not favor the majority class.
4. Hyperparameter Tuning: Fine-tune hyperparameters to better capture the underlying patterns in the data.
5. Threshold Adjustment: Experiment with different classification thresholds to improve precision and recall.