Pancreatic Cancer Detection using Radiomics: A Comparative Study between CNN Architecture and Vision Transformer

Abstract:

A CT scan or Computed Tomography scan image is a special type of medical diagnostic image that consists of a number of different Xray pictures taken from different directions around the body and then combined together by a computer to show a "slice" or cross-section of the region which is present inside the body. CT scans are usually painless, quick, comprehensive and they give valuable imaging information from multiple angles which can be helpful to doctors in making a comprehensive diagnosis..

Thus we can say that the CT-scan images help the medical experts to treat the patient and track the patient's health conditions by correctly identifying the nature of the disease. However, addressing the dynamic needs of a massive country like India, with a high clinical volume needs an alternative to manual human interpretation of diagnostic imaging . The manual process of analysing the CT scan image and determining whether there is cancer or not is a very tedious task that consumes a lot of time with a possibility that a mistake may occur in the process of predicting whether it was cancerous or not. Thus here for our work in order to simplify the process a bit we propose a pipeline for building a model based on neural networks which will try to accurately identify from the CT-scan image whether the image is cancerous or not. At first we build a CNN based architecture for image classification and later created a Vision transformer based architecture and found out that Vision transformer based models tend to perform better than CNN networks when it comes to image classification.

Introduction

Pancreatic cancer begins as an uncontrolled proliferation of cells in the pancreas, a retroperitoneal gland that has both exocrine and endocrine functions, being essential for digestion and blood sugar regulation. Unfortunately, this type of cancer is notorious for its high mortality rate due to the challenges associated with early detection. Symptoms in the initial stages are often vague and nonspecific, allowing the disease to progress unnoticed until it reaches an advanced stage, where treatment becomes significantly more difficult.

A **Computed Tomography (CT) scan** is a primary diagnostic tool for detecting and evaluating pancreatic cancer. These scans produce detailed cross-sectional images of the abdomen, enabling doctors to identify abnormalities in the pancreas, such as tumors or lesions, and to assess whether the cancer has spread.

The CT Scan Process:

Preparation: Patients may need to fast for several hours before the scan. Contrast dye, administered orally or via IV, is often used to enhance the visibility of organs and tissues.

During the scan: The patient lies on a table that moves through a circular scanner, which captures multiple X-ray images from various angles. These images are combined by computer software to create a comprehensive view of the pancreas and surrounding areas. The procedure is quick and painless, usually lasting 10 to 30 minutes.

Post scan analysis: Radiologists carefully examine the images to identify any abnormalities, focusing on factors such as the size, shape, and presence of tumors. If suspicious areas are detected, further tests—such as MRI, endoscopic ultrasound, or biopsy—are conducted to confirm the diagnosis and determine the cancer's stage.

Both CNNs and ViTs offer valuable tools for detecting pancreatic cancer from CT scans, each with strengths that complement traditional diagnostic approaches. CNNs are excellent for detailed, hierarchical feature extraction, while ViTs provide added benefits for modeling complex spatial relationships in medical images. Together, these AI technologies can play a critical role in improving early cancer detection, supporting radiologists, and ultimately increasing survival rates for pancreatic cancer patients.

The Aims and Objectives

The aim of this research is to create a pancreatic cancer detection system by leveraging the Neural Networks and comparing the CNN architecture and Vision Transformer architecture to find out which neural network among the two works best for Image classification for our project. The research objectives are as follows:

a) To create a CNN based architecture for Image classification to predict from CT-scan images whether there is cancer or not.

b) To create a Vision Transformer based architecture for Image classification to predict from CT-scan images whether there is cancer or not.

c) Conducting a comprehensive analysis to find out among the CNN architecture and Vision Transformer architecture which one works best for our research work.

Literature Review

Song et al. carried out the classification of oral cavity cancer images for mobile health applications by applying Vision Transformer and Swin Transformer models, demonstrating the effectiveness of these techniques for mobile health applications.

Zhang et al. explored Multiple Instance Learning with Shuffle Instances for the classification of pancreatic cancer ROSE images, showing its potential to improve diagnostic accuracy in pancreatic cancer detection.

Liu, Wu, Chen, Tsai, Roth, Wu, Liao, and Wang developed a deep learning model to distinguish pancreatic cancer tissue from non-cancerous tissue, validating it across different racial groups in a retrospective study for enhanced diagnostic reliability.

Urbanowicz, Suri, Cui, Moore, Ruth, Stolzenberg-Solomon, and Lynch proposed a machine learning model for biomedical binary classification in pancreatic cancer studies. Their study focussed on a nested case-control approach to eliminate bias in data analysis.

Li, Chou, Sun, Qiao, Yuille, and Zhou introduced an innovative label-free tumor synthesis method for early detection and localization of pancreatic cancer, aiming to enable non-invasive screening options.

Wu, Fang, Wang, and Shen investigated the application of deep learning techniques to predict pancreatic diseases from fundus images, proposing a novel approach for early disease detection through eye imaging.

Zhang, Feng, Feng, Zhao, Lei, Ying, Yan, and He implemented a Shuffle Instances-based Vision Transformer for the classification of pancreatic cancer ROSE images, demonstrating improvements in diagnostic performance.

Nagagopiraju, Thriveni, Reddy, Nithin, and Priyanka conducted research on detecting pancreatic cancer using machine learning and deep learning techniques, contributing to the development of more accurate diagnostic models for pancreatic cancer detection.

Elaanba, Ridouani, and Hassouni developed a transformer-based model to generate radiology text reports from frontal and lateral chest X-ray images, advancing automated medical reporting techniques.

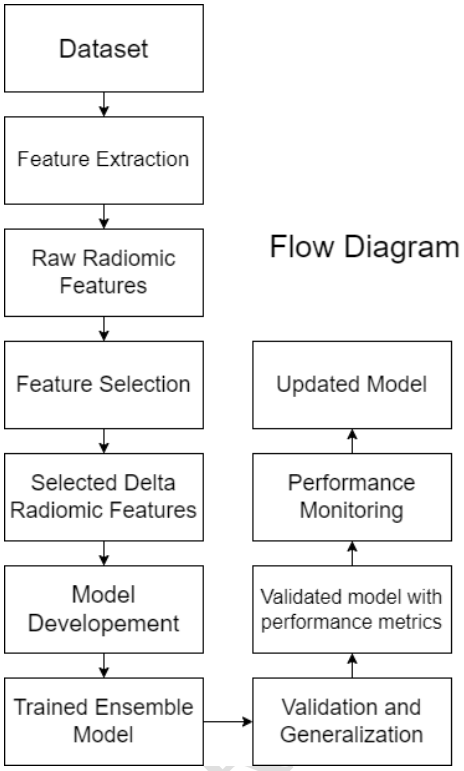
Slika, Dornaika, Merdji, and Hammoudi conducted research on a Vision Transformer-based model to quantify the severity of lung pneumonia from chest X-ray images, supporting more accurate assessment of pneumonia severity .

Song, KC, Yang, Li, Zhang, and Liang implemented Vision Transformer and Swin Transformer models to classify oral cancer images taken on mobile devices, demonstrating the potential of these models for mobile health diagnostics.

Ahmad, Alsulami, and Alqurashi applied transfer learning and Vision Transformers to enhance skin cancer detection, showcasing improved diagnostic performance through advanced deep learning techniques.

Ayana, Barki, and Choe used enhanced Vision Transformers to support early detection of colorectal cancer, emphasizing the model's potential for identifying pathological features at early stages.

Materials and Methods



Flowchart

Convolutional Neural Network

A **Convolutional Neural Network** (CNN) is a type of deep learning model particularly well-suited for image processing and classification tasks. CNN is powerful because it can automatically learn to extract important features from images through layers of convolution operations, without requiring manual feature engineering. Here is a detailed look at how CNN work while focusing on their application in image classification.

CNN Structure and Components

A CNN typically has several types of layers, each performing a specific function:

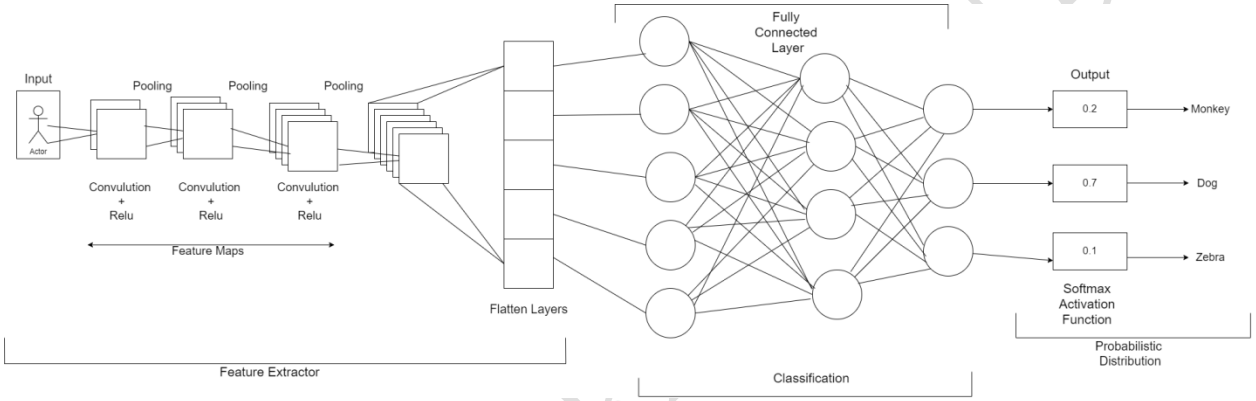
- **Input Layer**: This layer receives the raw pixel data of an image. Each image is represented as a matrix of pixel values, often 3-dimensional (height, width, color channels for RGB images).
- **Convolutional Layer**: This is the core layer of a CNN where the main computations happen. It applies filters (or kernels) to the input image to extract various features.

- **Pooling Layer**: Pooling reduces the spatial dimensions (height and width) of the feature maps. This helps make the network invariant to minor translations in the image and reduces the computational load.

- **Fully Connected Layer (FC Layer)**: After several convolutional and pooling layers, the CNN transitions to fully connected layers. These layers are similar to the layers in a traditional neural network and combine all extracted features to make the final prediction.

- **Output Layer**: The final layer of the CNN, typically a softmax layer for classification, provides the probabilities for each class.

Let's walk through these layers in more detail.



CNN Architecture

How a CNN Processes an Image

a.) Convolutional Layer

The **convolutional layer** is the core part of a CNN. It comprises of small, learnable filters (or kernels) that are layered onto the input image to create a **feature map**. Each filter detects specific patterns, such as edges, textures, or shapes, by scanning across the entire image.

1. **Filters**: A filter is a small matrix of weights , trained to identify certain specified features in the image. For example, an edge-detecting filter would highlight the borders of objects within the image.

1. **Strides**: Stride is the number of pixels the filter moves each time it slides over the image. A stride of 1 means the filter moves one pixel at a time; a larger stride would mean a faster but more sparse traversal.

2. **Padding**: Padding is adding extra pixels around the edges of an image to allow filters to apply even at the borders. This helps maintain the original image size after convolution.

3. **Activation Function**: ReLU (Rectified Linear Unit), an activation function is commonly applied to introduce non-linearity after convolution. ReLU helps the network learn complex patterns in the input image, by converting all the negative functions to a set zero value.

The output from each filter is a **feature map** that highlights different aspects of the image. For example, one filter may capture vertical edges, another may capture horizontal edges, and so on.

b.) Pooling Layer

The **pooling layer** is used to reduce the spatial dimensions of each feature map and control overfitting. The most common type is **max pooling**, which takes the maximum value in each region of the feature map.

1. **Max Pooling**: In max pooling, the feature map is divided into smaller regions, and only the maximum value from each region is taken. For example, in a 2x2 max pooling operation, the feature map is divided into 2x2 sections, and the highest pixel value from each section is selected to create a new, smaller feature map.

1. **Average Pooling**: Instead of taking the maximum value, average pooling calculates the average value in each region. This method can be used, but max pooling is more common in practice for feature selection.

Pooling reduces the size of each feature map, allowing the CNN to focus on important parts of the image while discarding less important details.

c.) Fully Connected Layer

After several convolutional and pooling layers, the network uses one or more **fully connected (FC) layers**. These layers connect every neuron from the previous layer to every neuron in the current layer, similar to a traditional neural network.

In the fully connected layers, the CNN interprets the high-level features extracted from the convolutional layers to make a decision on what the image represents.

d.) Output Layer

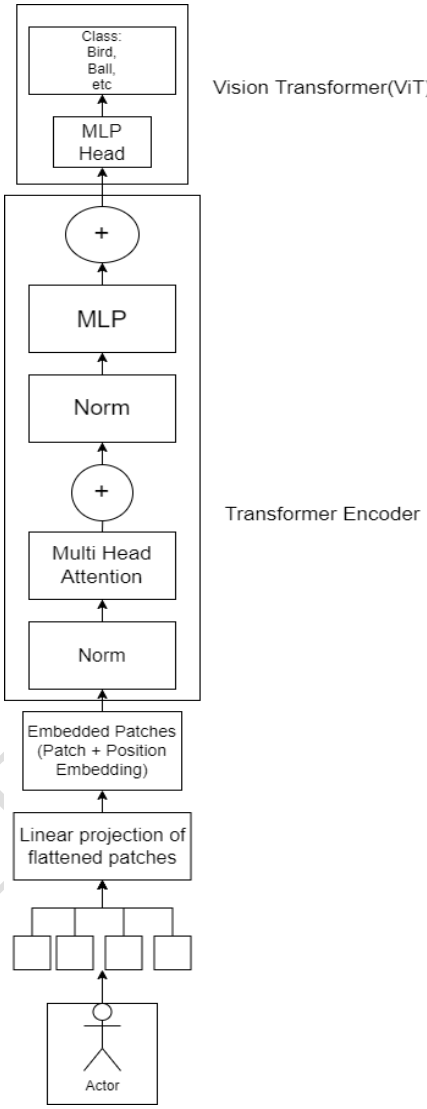
In an image classification task, the output layer typically uses a **softmax function** to produce probabilities for each class. The output neuron with the highest probability indicates the model's classification of the image.

For instance, in a CNN trained to recognize animals, if the softmax output for a given image shows 70% probability for "cat," 20% for "dog," and 10% for "rabbit," the CNN would classify the image as a cat.

Vision Transformer

Vision Transformers (ViTs) represent a newer approach to computer vision that departs from the conventional use of Convolutional Neural Networks (CNNs). While CNNs extract

features through convolutional filters, ViTs adopt the transformer framework originally introduced for natural language processing. By doing so, they treat an image as a sequence of patches rather than relying on convolution operations. This design shift has enabled ViTs to achieve strong performance on several computer vision benchmarks, making them a promising alternative to CNNs for image analysis.



Vision Transformer Architecture

A key component of Vision Transformers is the self-attention mechanism , which allows the model to assign different levels of importance to image patches during processing. This enables the network to capture contextual relationships and long-range dependencies within the image, improving its ability to represent complex visual patterns. In practice, self-attention generates attention scores between all token pairs in the sequence, producing an attention matrix that guides how information from each patch is aggregated .

The mechanism works by transforming each input embedding into three vectors—query, key, and value—through learnable linear projections during training. Attention scores are

calculated from the dot product of query and key vectors, then normalized with a softmax function to yield weights representing the relevance of each token. The output is a weighted combination of the value vectors, allowing every position to incorporate information from all other positions in the sequence .

To further improve representational power, ViTs use multi-head self-attention, where several attention heads operate in parallel, each focusing on different aspects of the image. Their outputs are concatenated and passed through a linear transformation, enhancing the model's ability to capture both global structures and fine details. The versatility of Vision Transformer's abilities across diverse computer vision task is attributable to this feature.. WQ, WK, WV are the learned weight matrices for the query (Q), key (K), and value (V) transformations.

$$Self\ Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right)V$$

$$Multi\ Head(Q,K,V) = Concat(head_1,\dots head_h)W^O$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

The multi-head self-attention block output is fed into a feed forward network (FFN) incorporating two linear activation functions and a rectified linear unit (ReLU) activation. X represents the output of the previous layer and the weight matrices of the first and second linear layers as Wa and Wb and the bias vectors as Ba and Bb.The generated output of the FFN can be calculated by the following formuala::

$$FFN = ReLU(XW_a + B_a)W_b + B_b$$

The Rectified Linear Unit (ReLU) is a widely used activation function that introduces non-linearity by applying an element-wise transformation to the inputs . By doing so, it enables the model to learn and represent complex, non-linear relationships at each position independently, which is essential for improving the expressive power of deep networks.

Experiments and Results

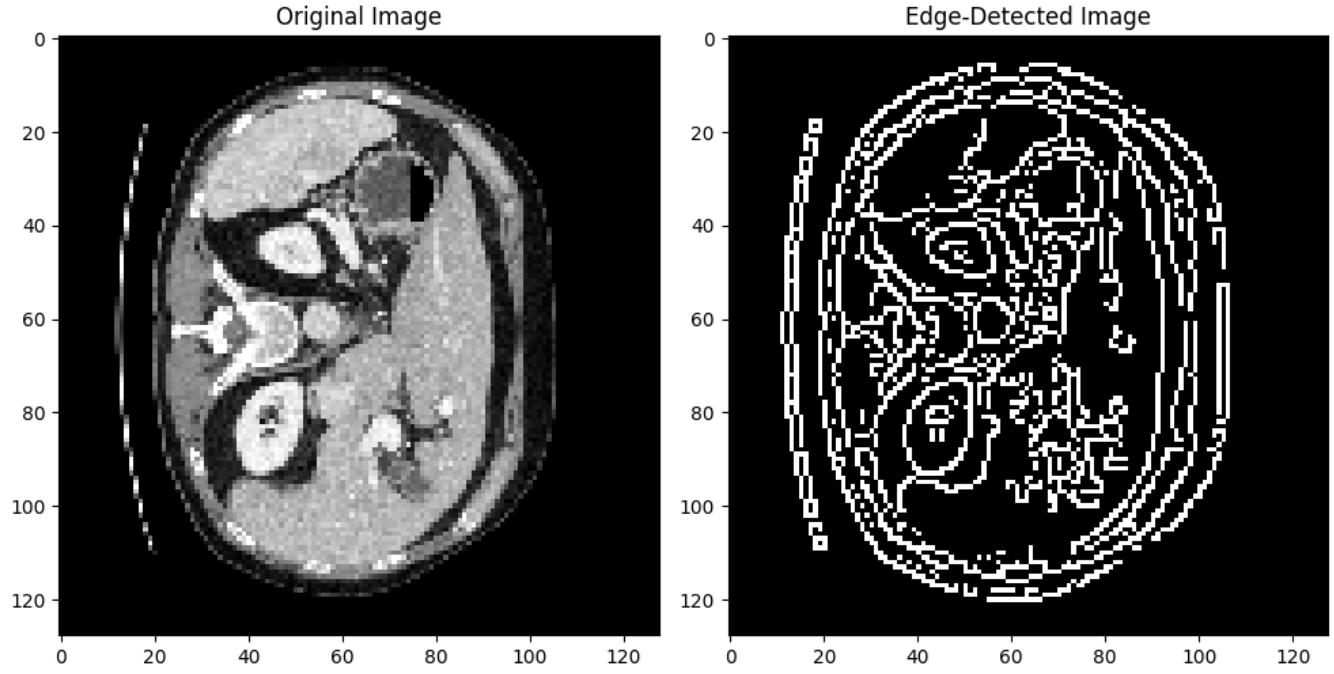
Dataset

The dataset employed in our study was obtained through Kaggle which is a part of a larger dataset taken from Cancer Imaging Program obtained through National Cancer Institute.

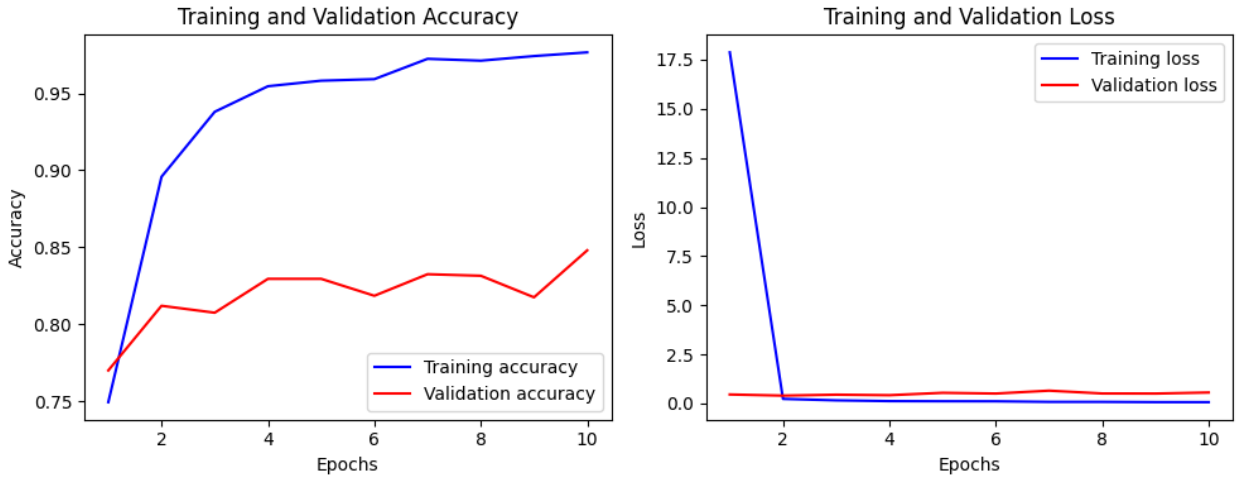
For our study we used a total of 14000 images. The images were seperated into 4 folders categorized as : a)X-Train folder which contains 5000 images of the CT scan images, b)Y-Train folder which contains 5000 images of the masks for the CT scan images which help in denoting which parts are cancerous in the CT scan images, c)X-Test folder which contains 2000 images of CT-scan images for testing and validation purposes, d)Y-Test folder which contains 2000 images of the masks of the CT scan images for testing and validation purposes.

The Experiments

In our study the VIT or Vision Transformer architecture is compared with a simple CNN architecture based on image classification for early detection of Pancreatic Cancer.The performance of these models were compared through epochs, batch size, validation data, prediction value and accuracy. We have 4 folders and in X-train and Y-train folders each containing 5000 images the images are pre processed such that their dimensions are changed to 128x128 and similarly in the X-test and Y-test folder each containing 2000 images the images are preprocessed such that their dimensions are changed to 128x128 as well.



Here we trained our model with 10 epochs and 32 batch size.



Finally we calculated the required parametric values starting with confusion matrix.

A **confusion matrix** is a table that breaks down how well a classification model is performing. It shows four key numbers:

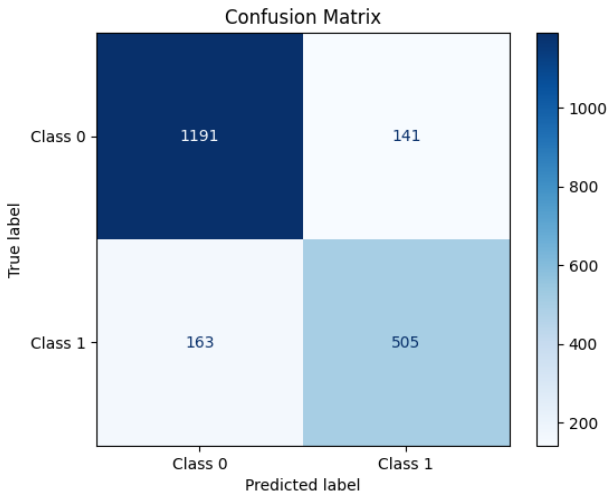
True Positives (TP): These are cases the model got right when it predicted something positive.

True Negatives (TN): Cases the model correctly identified as negative.

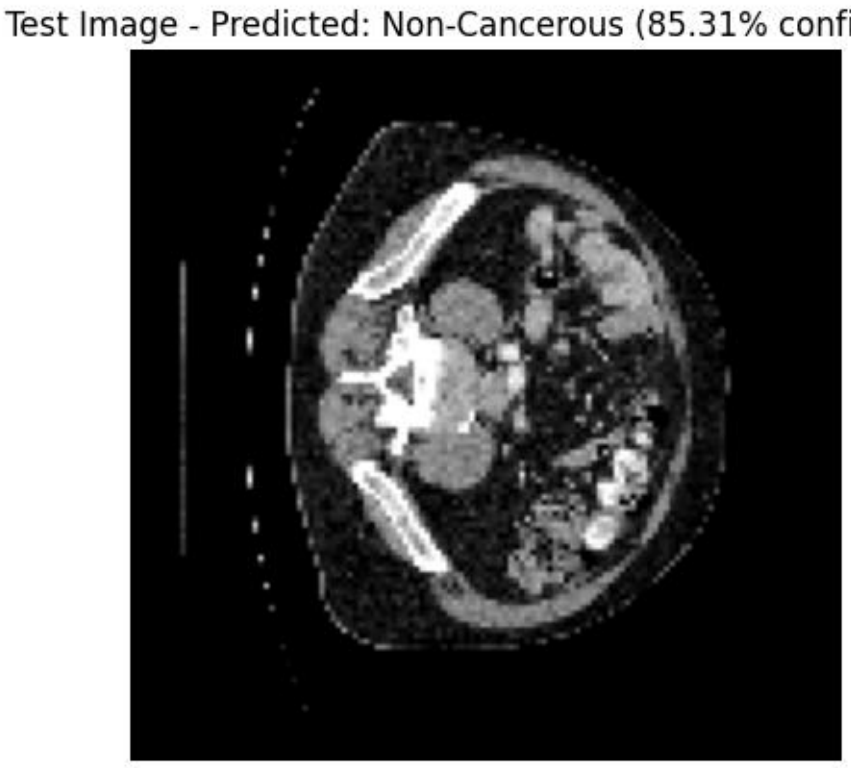
False Positives (FP): Times the model wrongly predicted something as positive when it wasn't.

False Negatives (FN): Times the model missed identifying something positive. By looking at this table, you can get a better sense of what your model is doing well and where it's making mistakes.

From our Confusion Matrix result here Class 0 refers to the Non cancerous class and Class 1 refers to the cancerous class



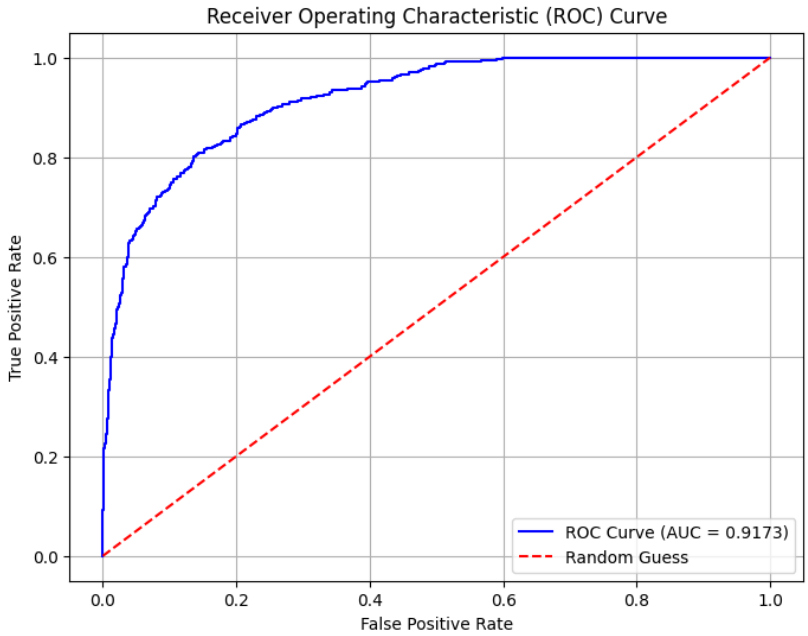
Accuracy: Accuracy is simply a way to measure how often our model is right overall. It's calculated by dividing the number of correct predictions (both positive and negative) by the total number of predictions. While accuracy can give us a quick idea of performance, it's not always the best measure—especially if our dataset is imbalanced, like for example when one class has far more samples than the other. The accuracy of my model came out to be 84.7 percent



Precision: Precision evaluates the accuracy of a model's positive predictions. It answers the question: "Among all the instances the model identified as positive, how many were actually correct?" High precision means the model makes fewer false alarms (i.e., it's good at not misclassifying negatives as positives). It's particularly important in scenarios where being wrong has serious consequences, like diagnosing a disease. The precision of the model is found to be 78 percent.

Recall: Recall (sometimes called sensitivity) is all about how well the model finds the things it's supposed to. It tells you what percentage of the actual positive cases the model correctly identified. A high recall means the model is good at catching positives, but it might not care much about false alarms. Recall of the model is found out to be 75.60

Area under curve(AUC): The Area Under the Curve (AUC) is a way to see how good a classification model is at telling two groups apart, like positive and negative cases. To figure this out, we use a graph called the ROC curve, which shows the trade-off between the true positive rate (recall) and the false positive rate at different thresholds. The AUC score is simply the size of the area under this curve.



The model accuracy of CNN came out to be 78 percent which means 78 percent of the times the model will give accurate result and rest 22 percent of the times the model may give inaccurate results.

However upon testing the model it was seen that the model was able to predict from CT scan images whether the image was cancerous or not with almost 80 percent confidence.

Similarly we implemented the same model using Vision transformer architecture and found out that the accuracy of the model using the Vision Transformers (ViT) architecture was better than Convolutional Neural Networks (CNNs) architecture for the same image classification in our project for the Early Detection of the Pancreatic Cancer.

Discussions

From our observations we see that Vision Transformers (ViT) performed better than Convolutional Neural Networks (CNNs) for image classification when the dataset is large and the application requires a high-level understanding of visual content. Although transformer-based models demonstrate strong performance, they also come with certain drawbacks . One major issue is their limited interpretability—while CNNs are already difficult to explain, transformers can be even more opaque. This lack of transparency poses challenges in medical settings, where clear explanations of model decisions are necessary to build trust with clinicians and patients . In addition, transformers generally demand higher computational resources than CNNs because of their self-attention mechanism and large number of parameters. As a result, deploying them in real-time or resource-constrained environments can be difficult. Hence, CNNs are still commonly used for smaller datasets and real-time applications because they are more efficient.