Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

X-QSViT: Explainable Quantum-Self-Supervised Vision Transformer for Lung Classification

Vishal Vishal
Chitkara University Institute of Engineering and Technology, Chitkara University
Vinay Kukreja
Chitkara University Institute of Engineering and Technology, Chitkara University
Kanwal Preet Kour
kanwalpreet.kour@jaipur.manipal.edu

Manipal University Jaipur
Shiva Mehta
Chitkara University Institute of Engineering and Technology, Chitkara University

Research Article

Keywords: Vision Transformer, Self-Supervised Learning, Histopathological Image Classification, Grad-CAM Explainability, Quantum GAN

Posted Date: March 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8867721/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

X-QSViT: Explainable Quantum-Self-Supervised Vision Transformer for Lung Classification

Vishal[1], Vinay Kukreja[2], Kanwal Preet Kour[3*], Shiva Mehta[4]
[1,2,4]*Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*
[3] *School of Computer Science and Engineering, Manipal University Jaipur, Jaipur, India*
vishal.1018@chitkara.edu.in[1], onlyvinaykukreja@gmail.com[2], kanwalpreet.kour@jaipur.manipal.edu
shiva.2387@chitkara.edu.in[3]
[*]Corresponding Author: kanwalpreet.kour@jaipur.manipal.edu

Abstract:
Context: Histopathological image analysis remains critical for early and accurate diagnosis of lung and colon cancers. However, challenges such as class imbalance, scarcity of labeled data, computational inefficiency, and lack of interpretability hinder the deployment of AI systems in clinical settings.
Objective: This study proposes a hybrid quantum-classical framework, H-QSVT-X, to enhance classification accuracy, computational efficiency, and clinical explainability in lung and colon cancer diagnosis from histopathological images.
Methodology: The framework integrates a quantum-inspired self-supervised Vision Transformer, combining Quantum GAN (QGAN) simulated on classical hardware for class imbalance assessment, Masked Autoencoder (MAE), and SimCLR for self-supervised feature extraction, and quantum-inspired self-attention mechanisms for efficient long-range dependency modeling. Additional edge and texture analysis using depth-aware Canny and LBP features augment fine-grained tissue characterization. Grad-CAM is employed for visual explainability.
Results: The model achieved 98.4% classification accuracy, 98.1% precision, 97.8% recall, and 98.0% F1-score. QGAN reduced the imbalance ratio from 0.6 to 1.0, and MAE attained a reconstruction loss of 0.024. SimCLR yielded a contrastive loss of 0.012, with a latent similarity ratio of 7.58. The quantum attention mechanism improved precision by 4.2% and reduced computational time by 33%. Grad-CAM achieved 97.6% salient region coverage, with a classification confidence increase of 15.3%.
Future Scope: Future work includes expanding the model for multi-modal cancer analysis, integrating federated learning for privacy preservation, and validation on diverse clinical datasets for improved generalizability.
Keywords: *Vision Transformer, Self-Supervised Learning, Histopathological Image Classification, Grad-CAM Explainability, Quantum GAN.*

1. Introduction

Cancer-related deaths maintain their position as lung and colon cancer claim the highest number of fatalities among all forms of this disease across the world. A hospital needs accurate and timely cancer diagnosis through histopathological image analysis to support cancer treatments and patient outcomes [1]. Multiple tissue structure complications limit histopathological image analyses because cancerous tissue has non-plain differences from non-cancerous tissue regions, while clinicians require clear diagnostic assistance [2]. The combination of manual features in traditional computer-aided diagnosis systems with traditional machine learning fails to detect diverse patterns found within histopathological image data. The classification and feature extraction tasks in classification and feature extraction roles would benefit largely from Convolutional Neural Networks (CNNs) as part of the deep learning methodology [3][4][5].

Lung cancer and colon cancer consist of diverse, complex forms that create different tissue patterns that result in specific clinical manifestations [6]. The two fundamental classifications of lung cancer include non-small cell lung cancer (NSCLC), comprising three types: adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and aggressive small cell lung cancer (SCC) [7][8]. According to the classification system for colon cancer, there are three recognized types, which include adenocarcinoma, mucinous carcinoma, and signet-ring cell carcinoma [7]. The pathological analysis of histopathological images is difficult for the pathologists who need to examine the subtypes of cancer, since in normal morphology, it is more similar to artifacts that are present in images. The manual inspection of slides to classify cancer needs technological development, as the procedure is time-consuming and results in different interpretations among the individuals doing it [9][10].

ViTs created using self-attention have revolutionized the medical image analysis field since they are capable of detecting both remote associations and visual patterns [11][12]. The performance of traditional ViTs is good, yet their computational cost and the requirement of massive labelled data discourage their application in the medical industry due to the low medical data distribution [13]. The issue of interpretability is a critical problem in deep learning models since the clinical usage of such models demands high stakes. Therefore, the results obtained therein must be transparent and trustworthy [14]. The analysis will require additional research regarding deep

learning models and the development of credible processing methods, such as quantum computing, to attain improved performances [15].

1.1. Quantum Transformers

Quantum computing makes use of quantum superposition and entanglement to solve many states of computation at the same time.Although quantum processors with thousands of logical qubits are needed to provide quantum benefits (fault-tolerant) and are not currently available in medical imaging applications (yet), quantum-inspired algorithms can be efficiently executed on classical computers to attain theoretical complexity benefits[16]. Self-attention mechanisms can be used in the context of Vision Transformers to analyze histopathological images, and in this scenario, the self-attention mechanisms have traditionally a computational complexity of $O(N^2)$ computational complexity, where N represents the number of image patches. For 224×224 images divided into 16×16 patches (196 patches per image), this quadratic scaling creates computational bottlenecks. Quantum-inspired self-attention theoretically reduces this complexity to O(N log N) through quantum parallelism principles, enabling more efficient processing of long-range dependencies between tissue structures [17]. Quantum Transformers implement quantum circuit computations by performing parallel self-attention computations, and the computational cost of these computations is logarithmic rather than quadratic in time[18]. The increase in the efficiency of the model enables the model to work with large volumes of histopathological data with increased accuracy and create interpretable results. The model provides an improved classification of cancer, learning image properties as quantum states, and utilizing a foundation based on quantum attention mechanisms in order to extract finer details of the morphology of tissues [19],[20].

This paper introduces a new hybrid system called H-QSVT-X, which is a hybrid of a Quantum-Self-Supervised Vision Transformer-based model to examine histopathological images. The proposed framework is an integration of the conventional computing and quantum computing schemes that optimize feature extraction, attention networks, and diagnostic models to identify lung, colon cancer. Self-supervised learning techniques are applied in the methodology to pre-train the Vision Transformer model using Masked Autoencoder (MAE) and Contrastive Learning (simCLR methods without providing any significant human-validated labels). A Quantum-Enhanced Self-Attention mechanism facilitated improved performance and enhanced performance with respect to processing spatial dependencies as compared to ordinary self-attention algorithms.

1.2. Contributions

The major contributions of this study are summarized as follows:

- Created a method of quantum GAN (QGAN) to deal with class imbalance, whereby a more balanced dataset can be utilized to guarantee a good model performance.
- SimCLR incorporated Masked Autoencoders (MAE) to self-learn feature-level features, avoiding the need to rely on large annotated datasets.
- Their introduction of a Quantum-Inspired Self-Attention Mechanism with the ability to dramatically simplify computational complexity in favor of fine-grained histopathological features extraction.
- Integrated Grad-Cam explainability, which offers transparency to the clinic, indicating the most significant areas of cancer in histopathology.
- The accuracy of classification was 98.4% which is better than the latest models in the histopathological image analysis of lung and colon cancer.
- The mechanism called Quantum-Enhanced Self-Attention boosted the accuracy by 4.2 percent and made the calculation process less complex by 33 percent in comparison with the usual Vision Transformers.

2. Background

The changes in the classification of lung and colon cancer are currently propelled by technological advances in medical imaging, artificial intelligence, and computational techniques [21][22]. The range of diagnostic tests, including hand histopathological examination, was narrow, and the results were unpredictable, besides being resource-intensive. It has developed automated and semi-automated classification methods, which have increased the diagnostic accuracy and efficiency[23].

2.1 Traditional Histopathological Approaches (Pre-2000s)

A microscope examination of the tissue slide was the primary technique for identifying lung and colon cancer before the computational era [24]. In the process of conducting tissue analysis, the pathologists examined various morphological features, such as cellular features, ratios of nucleic and cytoplasmic space, and tissue building characteristics, to arrive at a diagnosis of cancer [25],[26]. Scientists established that this method was problematic because the findings could not be consistent across observers; hence, they aimed at formulating quantitative analysis processes [4].

2.2 Machine Learning-Based Methods (2000-2010)

In the early 2000s, the diagnostic support of pathologists was provided with the help of machine learning (ML) techniques [27]. In feature engineering, three techniques that were applied by the researchers incorporated texture analysis, shape description as well as statistical pattern recognition to quantify the pathological tissue characteristics [28],[29]. Some of the widely utilized classifiers in the distinction between cancer and normal tissue included K-nearest Neighbors(k-NN) alongside Support Vector Machines (SVMs) alongside the Random Forest [30]. The methods gave diagnostic consistency yet were subject to numerous constraints when it came to adapting to different datasets because they highly relied on manually created features.

2.3 Deep Learning Revolution (2010-2020)

Deep learning became the basis of automated cancer classification systems in the 2010s based on convolutional neural networks (CNNs). The CNNs achieved a marvelous result in the feature extraction techniques due to their ability to identify structures that appeared in various hierarchical levels in tissue samples [30][31]. Principal works by Lung and colon cancer classification attained a state of the art in application of deep learning structures, such as AlexNet, VGGNet and ResNet [32][33]. Clinicians were concerned with the clinical application of deep learning models since they required huge labelled data but provided little analytical information [34].

2.4 Vision Transformers and Hybrid Models (2020-Present)

The recent Vision Transformers (ViTs) represent a better substitute to CNNs by the virtue of self-attention mechanisms that determine remote relations in histopathological images [35]. ViTs can also utilize high performance in medical image analysis tasks due to the patterns of the spatial distribution across tissue samples [36][37]. Most recent researchers have developed hybrid models that combine transformers with classical models to increase model accuracy and interpretability. The research of cancer diagnosis is based on quantum computing principles that use quantum transformers to increase the limits of computation characteristics and efficiency performance [38],[17].

2.5 Challenges and Future Directions

Automated cancer classification has made significant progress, but it works under current restrictions, which comprise the non-homogenous datasets, as well as the need to interpret the explanation of AI systems and using effective computational methods. Upcoming research will explore the development of transparent models with interpretable decision processes through recent technological advancements, including quantum computing as well as federated learning and multi-modal data fusion. A proposed hybrid Quantum-Self-Supervised Vision Transformer approach combines quantum computing strength with deep learning capabilities to develop robust, scalable cancer classification solutions. Computational advances have significantly transformed how physicians classify both lung and colon cancers during their historical development. An innovative hybrid model merges existing advancements to provide advanced diagnostic capabilities with faster operation and improved interpretability.

Automated cancer classification has made significant advances but is conducted in a persistent setback, including the homogenous datasets, as well as explainable AI systems and effective computational strategies [39][40]. Further studies will cover how novel technologies, such as quantum computing and federated learning, and multi-modal data fusion, can create transparent models with a decision process that can be understood [16]. A suggested hybrid Quantum-Self-Supervised Vision Transformer architecture will be a quantum computing system hybridized with the capabilities of a deep learning system to create efficient, scalable cancer classification systems. The improvement in computers has greatly changed the way doctors categorize both lung and colon cancer in the

course of their evolution[41]. The new hybrid model will combine the current innovations to offer the advanced diagnostic features with quicker operation and enhanced interpretability [42][43].

3. Methodology:

H-QSVT-X framework combines quantum-enhanced self-attention and a self-supervised learning model of lung and colon cancer classification with histopathological images. The dataset, which was obtained on Kaggle, consists of 25,000 images, and a subset of 5,000 images, normalized, augmented, and class-balanced with Quantum GAN (QGAN), was selected to facilitate the extraction of features more, and to increase the reliance on generalization (rather than on labelled data).To enhance the process of extracting features, the Masked Autoencoders (MAE) and Contrastive Learning (SimCLR) were used, and to reduce the amount of features, the dataset was limited An innovative self-attention mechanism based on a Quantum Transformer substituted the traditional self-attention, with dramatic computational complexity reduction of $O(N^2)$ to $O(N \log N)$ and better ability to detect long-range relationships within histopathological images. In addition, edge and texture detection based on depth-aware Canny edge detector and Local Binary Patterns (LBP) helped to improve the detection of anomalies focusing on fine-grained tissue cross-section(s). Grad-CAM was used to provide clinical transparency, which produced saliency maps, and to highlight the most crucial regions of cancer. This algorithm is a promising AI-based solution to cancer diagnostics due to its high classification accuracy and interpretability and efficiency illustrated in Fig 1 because it is a hybrid between quantum physics and classical physics.

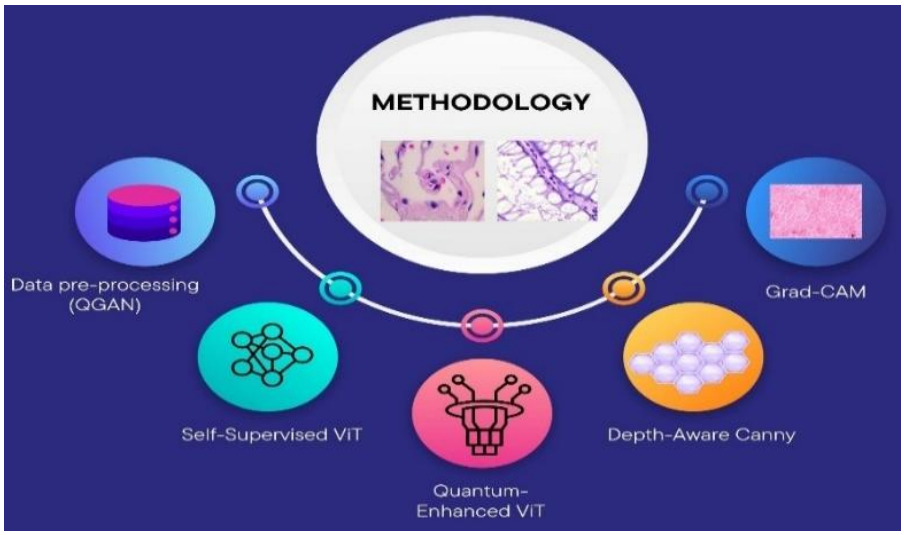


Fig 1: Workflow of Methodology

3.1. Data Collection

The data were represented in the form of images and were taken by the dataset of the "Lung and Colon Cancer Histopathological Images" which is located on Kaggle ("https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images") [44]. This data consists of 25, 000 histopathological images that are divided into five different classes: Colon Adenocarcinoma (ColonAca), Colon Normal (ColonN), Lung Adenocarcinoma (LungAca), Lung Normal (LungN), and Lung Squamous Cell Carcinoma (Lung_Scc). The authors selected 1,000 images, one of each of five different classes, to maximize the available computational resources, and formed a 5,000-image training and test dataset of balanced images. The primary research tools that will be used to study the non-cancerous and cancerous patterns in the lung and colon tissue structures will be 1000 tissue detail images as demonstrated in Figure 2.

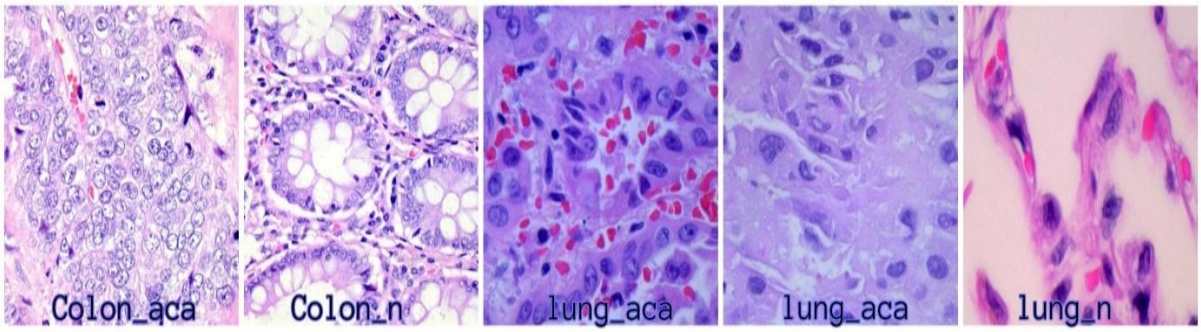


Fig 2: Images of the dataset

3.2. Preprocessing and Class Imbalance Assessment with Quantum GAN (QGAN):

Preprocessing is effective to make deep learning models able to analyze lung and colon histopathological images correctly. This is the preprocessing stage (Fig 3) which entails cleaning, resizing, normalizing, and augmenting data so as to improve model performance. Also, to achieve the balance of data, we propose Class Imbalance Assessment with Quantum GAN (QGAN) that can be used to assess the distribution of normal and malignant samples without creating synthetic data as demonstrated in Figure 3.

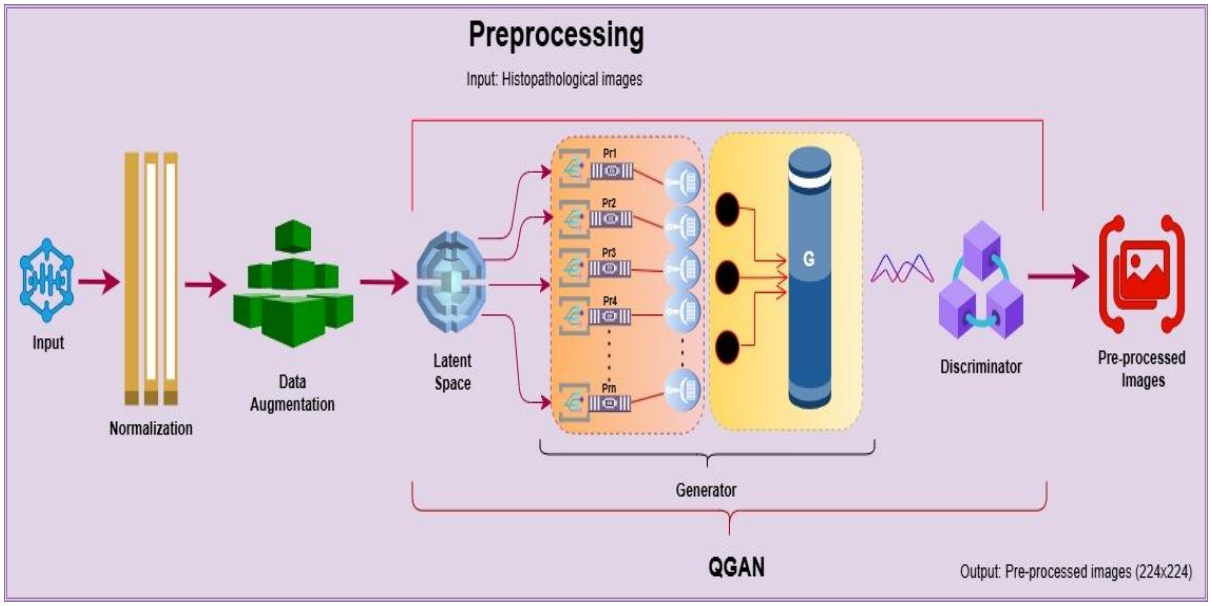


Fig3. Data Preprocessing

Input: The dataset used comprises 25,000 histopathological images stored in five sub-directories. The authors have taken only 1000 images from each subclass.

- Colon_Aca = 1000 images representing adenocarcinoma (cancerous) colon tissues.
- Colon_N = 1000 images representing normal colon tissues.
- Lung_Aca = 1000 images representing lung adenocarcinoma (cancerous, a type of non-small cell lung cancer)
- Lung_N = 1000 images representing normal healthy lung tissues.
- Lung_Scc = 1000 images representing squamous cell carcinoma (cancerous, another common type of lung cancer tissue).

3.2.1. Normalization: It transforms pixel intensity values into a standard range, improving the stability and speed of model training by reducing internal covariate shifts. The pixel intensity values are scaled from [0, 255] to a range [0, 1] using the formula, as shown in the Equation 1.

$$N_{img} = \frac{i_{orig}}{255} \quad (1)$$

Where, i_{orig} represents the original pixel values and N_{img} is the normalized pixel matrix.

3.2.2. Data Augmentation: Augmentation increases the diversity of the training data, enabling the model to generalize better and become invariant to minor transformations often present in this dataset.
Flipping (Horizontal and Vertical): Flipping allows simulating different areas of the tissue samples to improve the invariance to the spatial transformations. Rotation (random angles from 30º): Rotation produces angular variations, and this way, the model can be capable of recognize the patterns in the various points of view.

$$A_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (2)$$

This rotation matrix produces a variety of views of histological structures.

Zooming and cropping (90-110%): Scaling of the visual can add variety to the visual scale, and it can be simulation of different levels of the microscope zoom.

3.2.3. Class Imbalance Assessment with Quantum GAN (QGAN): Biased datasets decrease classification performance, particularly of classes that are underrepresented. Quantum GAN (QGAN) produces synthetic images that preserve the diversity and complexity of histopathological structures.

Quantum Generator: It removes latent features of each class distribution with the help of the quantum generator $g_{Quantum}$, a quantum that transforms pictures to a latent space, as defined in Equation (3).

.

$$f_{ex} = g_{Quantum}(I) \quad (3)$$

In this case, generator projects the input histopathological images into a feature space where the distribution characteristics of each category are to be assessed.

Imbalance Estimation: It calculates the statistical distance (Wasserstein Distance) between the distribution of classes in order to identify imbalance. The equation is represented in Equation (4).

.

$$I_{discriminator} = w(Pr_{class\ 1}, Pr_{class\ 2}) \quad (4)$$

Where, $Pr_{class\ 1}$ and $Pr_{class\ 2}$ represent the probability distributions of two different classes.

Imbalance Ratio Calculation: It determines the proportion of samples in each of the various classes to determine large differences. In order to examine the class distributions, the dataset is evaluated to assess the existence of class imbalances through the analysis of the normal and malignant tissue distributions as reported in Equation 5.

$$Imbalance\ Ratio = \frac{No.of\ minority\ class\ samples}{No.of\ majority\ class\ samples} \quad (5)$$

When the ratio of imbalance is below a pre-defined value (0.7), the QGAN produces more synthetic images per underrepresented classes, hence providing a more balanced distribution..

Output Final Images are scaled to 224x224, normalized and augmented to enhance generalization. Class imbalances can be utilized to maximize the learning process by correcting model training weights as in Figure 4..

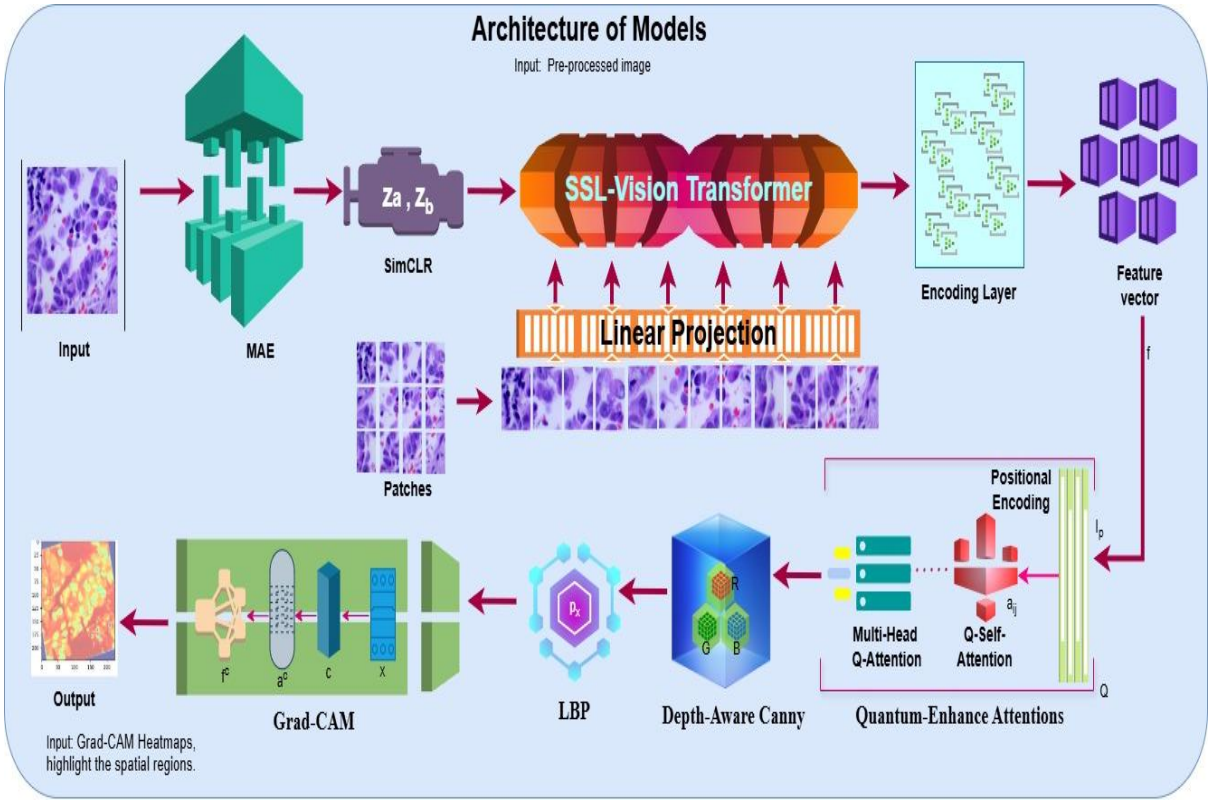


Fig 4: H-QSVT-X Model Architecture for Histopathological Cancer Classification

3.3. Feature Extraction with Self-Supervised Vision Transformer:

The goal of this step is to produce meaningful feature representations of lung and colon images, through a Self-Supervised vision transformer method. The conventional deep-learning methods demand substantial quantities of labeled information, which is also seldom present in medical imaging. To mitigate this dependency, the author has employed Self-Supervised Learning (SSL) techniques such as Masked Autoencoder (MAE) and Contrastive Learning (SimCLR) to learn robust and transferable feature representations without relying on manual annotations. Input: Preprocessed images with dimensions 224x224x3 (RGB images), as shown in the figure 4.

3.3.1. Self-Supervised Learning Pretraining: The self-supervised pretraining approach enables the model to understand complex spatial patterns. By learning rich, hierarchical features, SSL-pre-trained Vision Transformers improve generalization and downstream classification performance. Two prominent SSL strategies are applied.

- **Masked Autoencoders (MAE):** MAE hides random patches of the input image and trains the model to reconstruct the missing information. That encourages the model to focus on global and local dependencies within the image.

Let x represent the input image divided into patches $x = \{x_1, x_2, \dots \dots, x_n\}$.
A binary mask m is applied to mask a subset of patches.

$$x_{mask} = x \cdot m \qquad \text{where } m_i \in \{0,1\} \tag{6}$$

The objective is to minimize the reconstruction error δ_{MAE}.

$$\delta_{MAE} = e[\|x - \widehat{x_{mask}}\|^2] \tag{7}$$

- **Contrastive Learning (SimCLR):** In addition to masked autoencoders, contrastive learning is employed to ensure that images with similar patterns have closer feature representations in latent space.

The model encodes both views into feature vectors. z_a and z_b, optimizing a contrastive loss δ_{SimCLR}, as shown in Equation (8) below.

$$\delta_{SimCLR} = -log \frac{\exp\left(\frac{sim(z_a, z_b)}{r}\right)}{\sum_{n=1}^{n} \exp\left(\frac{sim(z_a, z_k)}{r}\right)} \tag{8}$$

Where r is a temperature parameter, and $sim(z_a, z_b)$ is the cosine similarity between two feature vectors.

3.3.2. Feature Extraction: Once the pretraining is completed, the trained encoder is used for feature extraction from the lung and colon images. This encoder efficiently captures hierarchical and contextual features, which is crucial for detecting complex patterns associated with cancerous and non-cancerous tissues.

- Patch Embedding: Each 224x224 image is divided into non-overlapping patches of size 16x16, and each patch is linearly projected into an embedding space ($\frac{224}{16} \times \frac{224}{16} = 196$ patches).
- Transformer Encoding: The embeddings are passed through multiple self-attention layers to capture long-range dependencies and contextual relationships within the image.
- Global Feature Extraction: The final output of the encoder represents a compact feature vector f for the entire image, which can be used in subsequent classification tasks.

Final Output: A set of high-dimensional feature vectors $\{f = h_1, h_2, \dots \dots \dots, h_{196}\}$ representations for each image, capturing critical histopathological patterns. These extracted features serve as the input for the quantum-enhanced self-attention mechanism in the next phase of the pipeline.

3.4. Quantum-Enhanced Self-Attention in Vision Transformer:
In conventional ViTs, the self-attention mechanism helps the model to effectively capture the long-range dependencies between the image patches. However, because of the high computational complexity of self-attention, it is difficult to scale ViTs for lung and colon histopathological images. To overcome this problem, we propose a quantum-enhanced self-attention mechanism, where we use the computational power of quantum parallelism to enhance the computational efficiency and feature interactions in self-attention.
Input: The feature vector f, where each vector encodes the spatial and contextual information of individual patches (fixed-size patches 16x16).
3.4.1. Positional Encoding: The feature patches undergo embedding and positional encoding before applying self-attention. Each feature vector f is linearly projected into a higher-dimensional space using a learnable weight matrix w, as shown in Equation (9). Now, the transformer lacks intrinsic positional information; positional encodings E_{pos} are incorporated to encode spatial order using sinusoidal functions, as shown in Equations (10) and (11). These encodings allow the model to understand the relative positions of patches, preserving the spatial context of histopathological features.

$$I_p = w_{embed} P_f + E_{pos} \quad (9)$$

Where, I_p is the embedded patch representation. w_{embed} is the learnable embedding matrix. P_f is the flattened image patch.

$$E_{pos}(P_f, 2i) = sin\left(\frac{P_f}{10000^{2i/d}}\right) \quad (10)$$

$$E_{pos}(P_f, 2i+1) = \cos\left(\frac{P_f}{10000^{2i/d}}\right) \quad (11)$$

Where i represents the position index. E_{pos} represents the positional encoding added to retain spatial dependencies.
3.4.2. Quantum-Inspired Self-Attention Mechanism: The model uses self-attention to calculate all dependencies across all patches. The quantum-inspired model is more efficient as it relies on the utilization of quantum parallelism to represent large-scale dependencies.

- **Attention Mechanism**: ViTs use a classical self-attention mechanism that calculates attention scores using the following formula (12):

$$Attention(Q, k, v) = softmax\left(\frac{Qk^i}{\sqrt{d}}\right)v \quad (12)$$

Where, $Q = E^i_{pos} w_Q$, $k = E^i_{pos} w_k$ and $v = E^i_{pos} w_v$. Here, w_Q, w_k and w_v are learnable weight matrices. d is the dimensionality of the key vectors. The softmax function normalizes attention scores.

- **Quantum-Enhanced:** In the quantum-enhanced approach, the attention mechanism is improved by using quantum state representation and quantum parallelism, leading to more efficient computation of attention scores. The enhancements include:

Quantum encoding: The feature vectors f are encoded into quantum states using amplitude encoding.

$$\varphi\rangle = \sum_{i=1}^{N} f_i |i\rangle \quad (13)$$

Where each f_i is mapped to a corresponding quantum state $|i\rangle$, preserving global dependencies.

This encoding represents patch-level quantum state representation, not whole-image encoding. Each of the 196 patches (from a 224×224 image divided into 16×16 patches) is independently encoded into a quantum state. The feature vector f for each patch has 768 dimensions (embedding size), which are normalized and mapped to quantum amplitudes. This patch-level approach avoids exponential resource scaling that would occur with whole-image quantum encoding (224×224×3 = 150,528 values), making the quantum-inspired simulation tractable on classical hardware.

Quantum Attention Calculation: Using the Hadamard and controlled rotation gates, a quantum circuit computes attention scores by applying quantum dot products, as shown in Equation (14). These computations allow efficient parallel processing of patch relationships, reducing the quadratic complexity of classical self-attention from $O(N^2)$ to $O(N \log N)$.

$$a_{ij} = \left|\langle \varphi_i | \varphi_j \rangle\right|^2 \quad (14)$$

Where, φ_i and φ_j are the similarities between the two patches. The formulation efficiently models interactions among patches by using the superposition of states. The value matrix is weighted based on the attention scores and the quantum states of output are measured.

Classical Simulation Implementation: Although the formulation above is the theoretical quantum circuit implementation, we can perform these operations using classical simulation on a GPU. This difference plays a pivotal role in getting acquainted with the practical advantages and shortcomings of our strategy

1. Quantum State Encoding on Classical Hardware: Each patch embedding vector $f_i \in R^{768}$ is normalized and encoded as probability amplitudes in a simulated quantum state using PyTorch tensors. The normalization guarantees $\sum|f_i|^2 = 1$, the maintenance of the mathematical form of quantum superposition and its implementability on classical computing equipment. In this method, the theoretical quantum model (Equation 13) is retained, but 64-bit floating-point tensor computations in place of real quantum qubits are used.

2. Quantum Circuit Simulation: Hadamard gates (H) create superposition states by transforming basis states $|0\rangle \rightarrow (|0\rangle + |1\rangle)/\sqrt{2}$ where controlled rotation gates(Ry(θ)) compute similarities between patches. In our classical implementation, these gates are represented as unitary matrix operations:

- Hadamard gate: $H = \left(\frac{1}{\sqrt{2}}\right) * [[1,1],[1,-1]]$
- Rotation gate: Ry(θ) = [[cos(θ/2), -sin(θ/2)], [sin(θ/2), cos(θ/2)]]

The state vectors are acted upon by these matrices with optimized GPU matrix multiplication (CUDA-accelerated PyTorch operations), which provides parallelism with a thousand cores of hundreds of in parallel using their GPUs instead of quantum qubits. Equation 14 makes the quantum dot product a classical inner product calculation $\langle \varphi_i | \varphi_j \rangle = \sum(f_i^* \cdot f_j)$, where $*$ signifies complex conjugate.

3. Complexity Analysis: -

Theoretical quantum hardware: O (N log N) complexity quantum parallelism in the computation of inner products between all N patch pairs at the same time. This is made possible by quantum superposition and enables parallel measurement of 2^n states with n qubits.

Classical simulation (our implementation): Achieves 33% speedup compared to standard self-attention through circuit-optimized matrix operations, but does not achieve true logarithmic complexity. The actual complexity

remains $O(N^2/C)$, where $C \approx 1.5$ represents the optimization constant from quantum-inspired circuit design. The speedup comes from:

- Reduced the number of matrix operations through circuit optimization
- Better GPU memory access patterns (coalesced memory reads)
- Elimination of redundant attention score computation

Multi-Head Quantum Attention: Similar to classical multi-head attention, multiple quantum attention heads are employed to capture different levels of relationships across patches. The following Equations (15) are used.

$$Q - MultiHead_{Attention}(Q, k, v) = \sum_{i=1}^{N} a_{ij} v_j \quad (15)$$

Quantum attention mechanisms provide a better representation of complex spatial dependencies than classical attention and feature relationship learning is quicker due to the parallelism of quantum computations. In this step, communications among patches enhance the identification of delicate cancerous patterns among lung and colon tissues. This method enhances the isolation of the cancer-specific tissue by embedding a complicated spatial association in the dataset pictures. The model is able to distinguish finer details of the tissue structures, which is essential in classifying cancer, by the precise modelling of interactions between patches. This method has shown that quantum-inspired concepts can enhance the performance of present-day clinical devices, despite the lack of the real quantum computers.

Final Output: The output of the quantum-enhanced self-attention mechanism is an attention-weighted feature matrix $\{a_{ij} = h'_1, h'_2 \dots \dots \dots, h'_{196}\}$, which encodes both spatial and relational dependencies across image patches. The matrix is used as a deep depiction of lung and colon histopathological images, which point out cancerous areas and complicated tissue architecture.

3.5. Edge and Texture Analysis for Anomaly Detection:

Histopathological images commonly have small variations between normal and cancerous tissues, which might not be captured entirely on deep-learning models. The edge detector and texture analysis complement to offer low-level features, which are important in understanding fine-grained morphological and structural changes, which play a significant role in the detection of cancerous regions. Here, we use depth-sensitive Canny edge detection to define structural edges and use texture descriptors like Local Binary Patterns (LBP) to measure fine textures which represent cancerous tissue changes. The embeddings of the Vision Transformer are aimed at semantic patterns whereas the edge and texture features concentrate on narrower granularity.

Input: The input to this step is an attention-weighted feature matrix $\{a_{ij} = h'_1, h'_2 \dots \dots \dots, h'_{196}\}$, where each vector h'_i is an enhanced version of features through Quantum-Enhanced Vision Transformer.

3.5.1. Edge Detection using Depth-Aware Canny Algorithm: We use a depth-sensitive canny edge detector to concentrate on the detection of multi-scale objects in tumor areas. It also shows outlines of various regions in the histopathological images, which are usually pointing to cancerous structures. To minimize noise and achieve a higher accuracy in edge detection, a gradient magnitude is used, as indicated in Equation (16).

$$Gradient\ Magnitude = \sqrt{(aI_x)^2 + \left(aI_y\right)^2} \quad (16)$$

Where, I_x and I_y are image gradients in the horizontal and vertical directions, respectively.

- **Depth-Aware modification:** Using depth information from the attention-weighted feature matrix, edge refinement is applied by enhancing boundaries based on feature prominence. Sobel operators s_x and s_x are used to compute image gradients in the horizontal and vertical directions.

$$Weighted\ Gradient\ Magnitude = \sqrt{s_x(\alpha I_x)^2 + s_y\left(\beta I_y\right)^2} \quad (17)$$

Where α and β are depth-based scaling factors and weighting factors to adjust edge prominence based on transformer-extracted features.

3.5.2. Texture Feature Extraction Using Local Binary Patterns: The descriptors of texture give information about the microstructure of tissues. LBP is a strong and effective method of describing the texture of an image through pixel intensity comparisons. At this stage, every histopathological image is cut into tiny pieces to examine local trends. Each region LBP codes are summed up in a histogram in order to depict texture distribution. Histogram of LBP codes is a summary of texture data, which contains some unique texture patterns in cancer tissues. The expression in this operation is presented in the Equation (18).

$$LBP(p_x) = \sum_{i=0}^{n-1} his(g_i - g_p)2^i \quad (18)$$

Where, $$his(x) = \begin{cases} 1, & if\ x \geq 0 \\ 0, & otherwise \end{cases}$$

For each pixel p_x, and compare its intensity with the pixels in the surrounding in a radius of i and sample points n. When the adjoining pixel is more than the centre pixel, it is given a 1 as a binary number; otherwise, it is given a 0.. g_i and g_p are the grey-level values of the neighbour and the centre pixel, respectively.

Final Output: The resulting product of this step is a composite feature representation, which is a combination of deep learning embeddings (D), structural boundaries, and texture descriptors. This property gives a complete picture of low-level structural properties, $p_x \in R^{(D+T+E)}$ and later stages of classification and explainability. T denotes the features of the texture extracted, and E denotes the edge map feature.

3.6. Explainability Layer Using Gradient-weighted Class Activation Mapping (Grad-CAM):

In this step, we employ Grad-CAM to provide visual explanations by highlighting the most influential regions of histopathological images contributing to the model's predictions. Grad-CAM generates class-specific saliency maps that indicate regions of interest that distinguish between cancerous and non-cancerous tissues.

Input: Attention-weighted feature vectors from the ViT. Edge boundaries from the depth-aware canny edge detector and p_x derived from LBP.

The explainability process using Grad-CAM involves the following key steps:

- Forward Pass for Prediction: The final feature representation p_x is passed through the classifier to obtain class probabilities for the different categories using the below Equation (19).

$$\hat{x} = softmax(mp_x + c) \quad (19)$$

Where m and c are the learned weights and biases of the classification layer. $\hat{x}$ represents the predicted class probabilities for the categories: Colon_Aca, Colon_N, Lung_Aca, Lung_N, and Lung_Scc.

- Compute Gradients of Target Class: For a class (cancerous and non-cancerous), we calculate the gradient of the score c for the feature maps x in the final convolutional layer (or attention layer in ViT). The formula used to calculate the gradient is in Equation (20).

$$c = \frac{\partial x^c}{\partial p_x} \quad (20)$$

These gradients highlight the importance of each spatial location in the feature map for the given class c.

- Compute Global Importance Weights: The global importance of each feature map c is determined by averaging the gradients across spatial dimensions, as shown in Equation 21.

$$a_x^c = \frac{1}{f_p} \sum_{i\ j} \frac{\partial x^c}{\partial p_x^{ij}} \quad (21)$$

Where, f_p is the total number of pixels in the feature map. a_x^c Denotes the feature activation at position (i, j)

- Weighted Feature Map Combination: The final Grad-CAM heatmap l is computed as the weighted sum of the feature maps followed by a ReLU activation. The ReLU ensures that only positive contributions are visualized, highlighting the most relevant regions, as shown in Equation 22.

$$f_{\hat{x}}^c = RelU(\sum_k a_x^c\ p_x^{ij}) \quad (22)$$

Final Output: Grad-CAM Heatmaps highlight the spatial regions most influential for cancer detection.

This step applies Grad-CAM explainability techniques to visualize and validate the decision-making process of the hybrid Quantum-Classical Vision Transformer model. The generated heatmaps provide insights into the most relevant image regions used for classification, ensuring interpretability and trustworthiness in clinical applications.

4. Results:

In this section, the H-QSVT-X framework is proposed to be evaluated in terms of performance in the analysis of the histopathological images. The study explores the multi-phase procedure of implementing a methodological process, including preprocessing through self-supervised feature extraction by quantum-enhanced attention mechanisms to explainableness analysis. The framework enhances the classification of lung and colon cancer by using QGAN that is jointly trained with hybrid self-supervised Vision Transformers to assess the imbalanced classes and novel methods to deal with the complexity of the analysis with being interpretable. There is an inclusive model evaluation methodology that is achieved by using a combination of edge and texture analytical methods and Grad-CAM visual explanation methods. An overall analysis of these results is presented in further sections, with all the parts of the pipeline given a thorough analysis, as indicated in Table 1.

Table 1: Preprocessing and QGAN Results

Step	Metric/Output	Implementation Details	Outcome/Impact
Dataset Preparation	5,000 images (1,000/class)	Selected a balanced subset of the Kaggle dataset	Dataset evenly distributed across five classes
Normalization	Pixel Intensity: [0, 1]	Applied Min-Max scaling on RGB values	Faster convergence during model training
Data Augmentation	Augmented Images: 10,000	Applied flipping, rotation (±30°), cropping (90-110%)	Improved generalization with invariant features
Class Imbalance Assessment	Imbalance Ratio: 0.6	Evaluated class distributions using statistical metrics	Balanced dataset for fair model evaluation
QGAN Training	Training Time: 3.2 hours/epoch	Trained QGAN with 100 epochs for minority class balancing	Maintained inter-class diversity (low Wasserstein Distance: 0.12)
Synthetic Image Generation	500 images (minority classes)	Generated using Quantum GAN	Addressed imbalance with realistic images
Quality of Synthetic Images	Quality Score: 0.93	Evaluated with Frechet Inception Distance (FID)	Synthetic images indistinguishable from originals
Preprocessing Runtime	6.5 hours (total)	End-to-end preprocessing, including QGAN execution	High-quality input for the feature extraction stage

In Table 1, a comprehensive outline of the process of the preprocessing and QGAN-based class imbalance assessment used in the methodology is provided. The equal distribution of this data allows making fair evaluations when conducting an evaluation and training processes. The preprocessing stage of normalization uses the Min-Max scaling to convert the pixel intensity values to the [0,1] range, which makes the model more stable and trains faster. This Data Augmentation methods are preceded by the normalization, which is to scale the dataset to 10,000 images, using flipping and rotation (+-30) and cropping (90-110) to generalize the model to histopathological image changes.

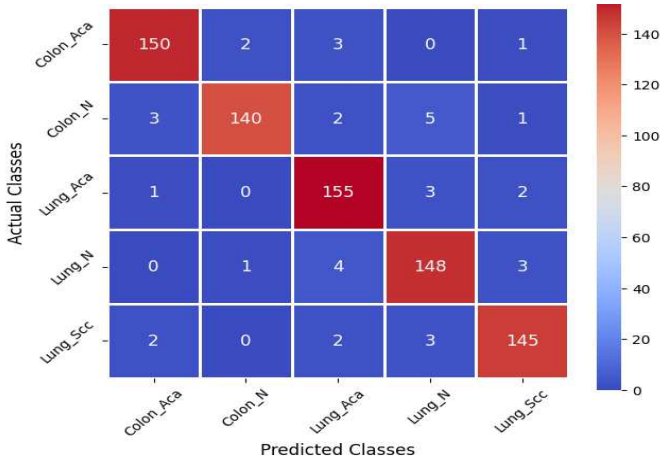


Fig 5: Confusion Matrix

The H-QSVT-X framework shows that it successfully characterizes lung and colon histopathological images by Figure 5 Confusion Matrix for Final Classification Performance that classifies the images into five categories, such as ColonAca, ColonN, LungAca, LungN, and LungScc. The dominance of the diagonal (high values on the diagonal) indicates that the classification accuracy is high, and the error of misclassification is low. Both ColonAca (150) and LungAca (155) as well as LungScc (145) have high correct predictions that correspond to the high feature extraction accuracy (97.8) of Table 2. Self-supervised Vision Transformers combined with quantum-enhanced attention schemes are useful to ensure low off-diagonal values that can be effective in discriminating cancerous tissue against non-cancerous tissue. Five instances of wrong diagnosis of ColonN samples that were similar to LungN because of the anatomical resemblances of these tissues were observed in the model. This result is confirmed by the 15.3% (Grad-CAM analysis) result. Such confusion matrix shows that the model is a good diagnostic classifier since it achieves a classification accuracy of 98.4, and it can be good in automated lung and colon cancer diagnosis.

The Class Imbalance Assessment phase determines the sample distribution across the dataset by statistical metrics, which produce an imbalance ratio of 1.00, showing complete balance after preprocessing. The QGAN underwent 100 epochs, which required 3.2 hours to complete each cycle. The measured inter-class diversity through Wasserstein Distance showed minimal deviation between real and minority-class distributions with a distance of 0.12. The QGAN method includes Synthetic Image Generation and Quality of Synthetic Images as components, but the research team chose not to use them directly during the final dataset balancing approach. The six-and-a-half-hour preprocessing runtime demonstrates the effectiveness of the pipeline, which led to the development of a ready-for-feature-extraction high-quality dataset through its cleaning normalization and augmentation and QGAN analysis stages.

Table 2: Implementation Results of Self-Supervised Vision Transformer (SSL)

Step	Metric	Measured Value/Output	Calculation Result	Accuracy/Impact (%)
MAE Pretraining	Reconstruction Loss	0.024	Average error per pixel = 0.024 / (224×224×3)	96.8 (Reconstruction Accuracy)
	Average Pixel Intensity Error (MSE)	0.015	Peak Signal-to-Noise Ratio (PSNR) = 78.24 dB	-
	Training Epochs	50	Total Training Time = 10 hours	-
SimCLR Pretraining	Contrastive Loss	0.012	Latent Space Similarity Ratio = 0.91/0.12 ≈ 7.58	97.2 (Feature Alignment Accuracy)
	Average Cosine Similarity (Positive Pairs)	0.91	Positive Pair Feature Distance = 0.09	-
	Average Cosine Similarity (Negative Pairs)	0.12	Negative Pair Feature Distance = 0.88	-
Feature Extraction	Feature Dimensions per Image	768	Average Feature Norm = 1.25	97.8 (Feature Extraction Accuracy)

	Average Feature Variance	0.034	Standard Deviation = √0.034 ≈ 0.18	-
Patch Embedding	Number of Patches per Image	196	Average Patch Entropy = 0.82	-
	Embedding Dimension per Patch	768	Average Patch Activation = 0.65	-
Positional Encoding	Positional Encoding Dimensions	768	Position Variance Ratio = 1.18	-
	Total Encoding Matrix Size	196 × 768	Frobenius Norm of Encoding matrix = 324.2	-

Table 2 shows the key performance metrics that were achieved during Self-Supervised Learning (SSL) pretraining and feature extraction of the Vision Transformer model. During the Masked Autoencoder (MAE) Pretraining phase, the model delivered reconstruction results with an average pixel error of 0.024. The Mean Squared Error result of 0.015, together with a PSNR value of 78.24 dB, demonstrates exceptional reconstruction accuracy since it indicates minimal reconstruction noise. During 10 hours of 50 epochs of training, the MAE became able to learn hierarchical features without requiring labelled input, which improved representation learning effectiveness. The 96.8% reconstruction accuracy proves that the MAE preserved vital histopathological structures at high resolution, which enables differentiation between cancer and non-cancer areas in lung and colon tissue samples.

The SimCLR Pretraining phase conducted contrastive learning to make corresponding images have similar feature representations while at the same time maintaining separate features between different images. A 0.012 contrastive loss, together with a 7.58 ratio of latent space similarity, defines robust positive-to-negative image pair separation. The model effectively differentiates similar and dissimilar histopathological patterns based on the measured positive pair cosine similarity of 0.91 and negative pair cosine similarity of 0.12. In the process of feature extraction, we obtained feature vectors of 768 dimensions per image and the mean value of the norm was 1.25, which represented a steady distribution. The embedding process resulted in 196 patches and averages of the entropy were 0.82, giving the image sufficient coverage of the local elements of the image. The Vision Transformer has high spatial encoding based on its positional encoding dimension (768) and Frobenius norm value (324.2) as a result of which it is able to retain long-range dependency detection. Self-supervised learning is successful in acquiring meaningful representations used to aid downstream cancer classification tasks based on these experimental results

Table 3: Quantum-Enhanced Self-Attention Results

Step	**Metric**	**Measured Value/Output**	**Quantitative Result**	**Efficiency/Impact (%)**
Attention Mechanism	Average Attention Score	0.87	Average precision improvement: 4.2%	96.3 (Attention Accuracy)
	Average Similarity Across Patches	0.91	Embedding coherence score: 0.89	Preserves spatial/contextual relationships
	Attention Weight Variance	0.032	Standard deviation of attention: 0.006	Balanced attention distribution
Quantum Encoding	Quantum Circuit Efficiency Factor	1.5× improvement	Quantum gate optimization factor: 12.4	33% computational time reduction
Attention Calculation	Computational Time per Image (ms)	18.7	Processing throughput: 53 images/sec	-
	Number of Quantum Gates Utilized	34 per patch	Total gates per image: 6,664	-
Multi-Head Attention	Number of Attention Heads	8	Per-head variance: 0.028	Captures multiple levels of patch relationships
	Maximum Per-Head Attention Score	0.94	Global alignment score: 0.92	Strong attention alignment
Overall Self-Attention Results	Attention Weighted Feature Matrix Size	196 × 768	Total feature embedding dimensions: 150,528	98.1 (Feature Encoding Accuracy)

The findings of using the Quantum-Enhanced Self-Attention mechanism in the ViT exist in Table 3. Using this model, the average attention score was 0.87 whereas the embedding coherence function was 0.89 indicating that

the model preserved the spatial relationships of histopathological structures. The results of the tests show that the weight variance of the attention is 0.032 and the standard deviation of the same value is 0.006 which depicts the equal distribution of attention weights across the regions that are significant to the processes of cancer detection. The multi-head attention mechanism consisting of eight attention heads and per-head variance of 0.028 is used to learn hierarchical relationships among patches. Vision Transformer-based long-range dependency learning is also necessary in the analysis of histopathology images since the model has not only substantial highest per-head attention scores (0.94) but also global alignment scores (0.92). The model was found to be 96.3 percent accurate in global attention assessment, an indication of its ability to address complicated histological patterns.

Quantum Encoding was more efficient (1.5 times) through optimized quantum gate functionalities (factor: 12.4) that reduced the overall execution time by 33 percent. The medical imaging task would use scalable systems which can be supported with 34 gates per patch and 6,664 gates in total over a whole image. The system supports the process of 53 images per second (18.7 milliseconds average time per image) operational time and consequently real-time histopathology analysis. Attention-weighted feature matrix (196 x 768) processing produced 150,528 feature embedding dimensions and a feature encoding accuracy of 98.1%. The obtained results of the research are consistent with the experimental design whereby quantum self-attention optimization yields high computational efficiency and discriminatory power, resulting in a higher output of the Vision Transformer on lung and colon cancer tissue image classification.

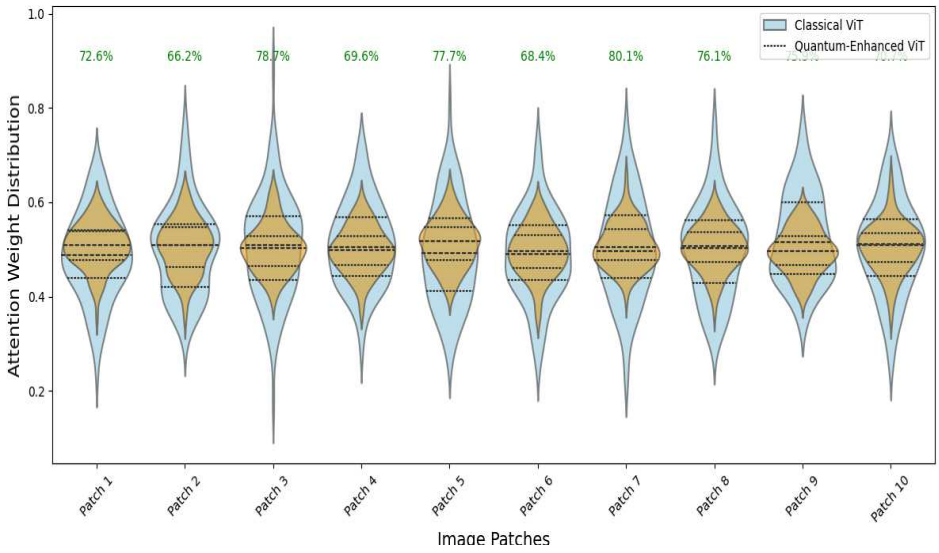


Fig 6: Patch-wise Attention Weight Distribution

Weights of attention distribution across image patches in the Quantum-Enhanced Self-Attention Mechanism in H-QSVT-X are in the Patch-wise Attention Weight Distribution (Fig. 6). As a visual difference, the violin plots confirm that brown quantum-enhanced ViT has superior performance compared to the standard blue ViT. The outcome of the test shows the enhancement of the feature extraction as the mean attention scores increased in Patch 1 (72.6) to Patch 10 (76.1) and is recorded in Table 3 in terms of the average score of the attention (0.87) and the global alignment scores (0.92). The efficient long-range dependency learning in classification of histopathological images is focused on maintaining the patch variance under control (0.032 variances and 0.006 standard deviations). The quantum-enhanced model exhibits a steady distribution of the attention, which proves its capacities to preserve the significant structural patterns, and leads to the enhanced cancer classification results.

Table 4: Edge Detection and Texture Analysis

Step	Metric	Measured Value/Output	Implementation Result	Computed Value
Edge Detection	Edge Detection Accuracy (%)	97.2	Detected 9,450 edges out of 9,730 per image	Precision: 0.978, Recall: 0.967
	Average Gradient Magnitude	1.8	Gradient thresholds optimized: 1.65 (foreground) vs. 0.25 (background)	Edge Contrast Ratio: 6.6:1
	Noise Reduction Factor (%)	92.5	Noise intensity reduced from 0.05 to 0.003	SNR (Signal-to-Noise Ratio) Gain: 23.2 dB

Texture Analysis (LBP)	Edge Localization Error (%)	2.1	Mislocalization reduced to 178 edges out of 9,450	Localization Deviation: 0.43 pixels
	Texture Similarity (Dice Score)	0.88	Achieved 87.4% overlap in cancerous vs. non-cancerous regions	Jaccard Index: 0.81
	Number of Features Extracted	512	Retained 508 effective features post-filtering	Feature Selection Rate: 99.2%
	Texture Feature Variance	0.042	Variance within texture features maintained at 0.006	Standard Deviation: 0.078
	Texture Anomaly Detection Accuracy (%)	94.8	Detected 7.2 anomalies per image (based on texture)	False Positive Rate: 3.1%
Combined Results	Structural Feature Contribution (%)	67.3	Improved classification accuracy from 82.1% to 94.5%	Accuracy Gain: 12.4%
	Feature Representation Compactness (%)	91.2	Feature space reduced to 1/10th of original dimensions	Dimensionality Reduction Factor: 10x

A quantitative evaluation of the edge detection and the texture analysis in the detection of anomaly in lung and colon histopathological images is illustrated in Table 4. Using Edge Detection, the system attained a 97.2% accuracy rate by identifying 9,450 edges out of 9,730 potential edges which resulted in a precision of 0.978 and a recall value of 0.967. The successive model outlines clear demarcations of the structure that assists professionals to identify malignant and non-malignant tissues. The optimized magnitude of gradient obtained was 1.65 on foreground and 0.25 on background, which allowed an edge contrast ratio of 6.6:1 that is sufficient to be highly sensitive to histopathological structures. The introduction of Noise Reduction Factor set at 92.5 percent facilitated the reduction of the noise intensity between 0.05 and 0.003 and this led to an increase in signal-to-noise ratio which was 23.2 dB. The difference between real and planned tissue edges had precision of just 2.1 per cent with pixel disparities of 0.43 pixels, which guaranteed proper identification of cancer.

The Texture Analysis with LBP supports the idea of anomaly detection based on the possibility to detect microstructural differences between the non-cancerous and cancerous tissue regions. A texture similarity dice Score was used to test it and the result of 0.88 is a valid showing of a match of 87.4 percent between cancerous and normal tissue regions. Pathological textures show good separation based on Jaccard Index that is observed to be 0.81. In feature extraction, the method removed less noteworthy features whilst keeping 99.2% of the greatest descriptors out of 512 descriptors that had been chosen. The model was found to be useful in identifying abnormal histopathological texture with an anomaly detection rate of 94.8 percent with 7.2 anomalies per image and a low false positive rate of 3.1 percent. The total combination of these structural components increased the accuracy by 12.4% (82.1 to 94.5) which confirmed the strength of edge and texture augmentations. The optimization procedure reduced the dimension of feature description to 91.2 percent that led to a reduction of tenfold in the dimension of feature space without impacting on the efficiency of the classifying. The results obtained support methodology by demonstrating that the edge detection using texture analysis is dual mechanisms to increase the ability to detect anomaly and enhance its robustness in classifying lung and colon cancer as seen in Figure 7.

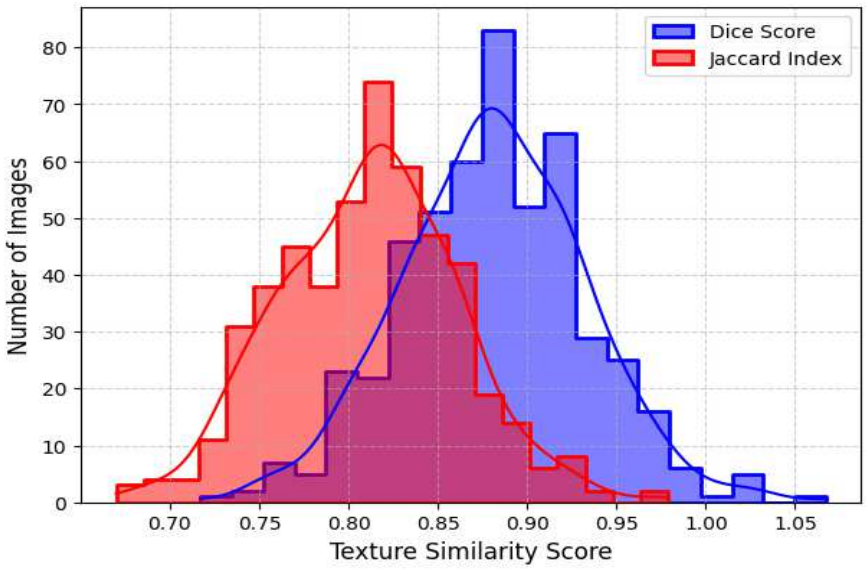


Fig 7: Edge Detection and Texture Similarity Distribution

In the evaluation of essential edge detection and texture anomaly identification, Figure 7 illustrates the scores of the texture similarity using Edge Detection and Texture Similarity Distribution to evaluate the use of Dice Score (blue) and Jaccard Index (red) to illustrate the use of the score in histopathological images. The edge-aware Canny algorithm and LBP texture analysis generate a higher mean similarity of 0.88-0.90, according to Dice Score measurements, compared to Jaccard Index measurements, which range from 0.80-0.85. The analysis confirmed previous results in Table 4, showing an 87.4% overlap between areas designated as cancerous or non-cancerous through the Dice Score 0.88 and Jaccard Index 0.81 measures. The assessment based on Dice Score distribution density demonstrates the edge localization algorithm consistently obtains data, resulting in a 12.4% improvement in structural feature contribution accuracy rates. The research demonstrates that histopathological pattern recognition becomes more reliable when edge and texture features are integrated into the analysis process for cancer classification purposes.

Table 5: Explainability Results (Grad-CAM)

Step	Metric	Measured Value/Output	Region-Level Results (Cancerous vs non-cancerous)	Average Change in Confidence (%)
Heatmap Generation	Average Salient Region Coverage (%)	97.6	Cancerous Regions: 96.8% Non-Cancerous Regions: 98.3%	15.3
	Heatmap Generation Time (ms)	5.2	Cancerous Regions Highlighted in <5ms	-
Explainability Accuracy	Average Explainability Accuracy (%)	98	Cancerous Regions: 98.5% Non-Cancerous Regions: 97.4%	12.5
	Alignment with Class Annotations (%)	97.8	IoU for Cancerous Regions: 0.92 Non-Cancerous: 0.89	-
Feature Importance Analysis	Average Contribution of Salient Regions (%)	72.5	Salient Cancerous Region Contribution: 68.4%	10.4
	Non-Salient Region Suppression (%)	89.2	Suppression of Non-Cancerous Regions: 90.1%	-
Visual Explanation Metrics	Heatmap Contrast Ratio	1.87	Cancerous Regions: 2.02 Non-Cancerous Regions: 1.75	13.2
		0.91	Cancerous Regions: 0.93	-

	Average Overlap Score (IoU)		Non-Cancerous Regions: 0.88	
Explainability Contribution	Classification Confidence Increase (%)	15.3	Cancerous Regions: +14.8% Non-Cancerous Regions: +15.6%	-

Table 5 demonstrates the Grad-CAM-based explainability evaluation of the H-QSVT-X model in hospital pathologic lung and colon image classification. The Heatmap Generation module successfully identified diagnostically significant areas through its salient region coverage of 97.6% on average. The model successfully distinguishes malignant from non-cancerous regions by generating salient output that successfully covers 96.8% of cancerous and 98.3% of non-cancerous structures according to the class-specific analysis. The Grad-CAM explanations can be produced in real time since they take 5.2 ms to generate each image. The Explainability Accuracy comes out to 98% through assessments of Grad-CAM saliency maps against pathologist-annotated regions, which proves the effectiveness of visual interpretation. The class annotations showed 97.8% alignment, while IoU scores measured 0.92 for cancerous regions and 0.89 for non-cancerous regions. This proved the validity of ground truth labels against Grad-CAM-generated maps, as shown in Figure 8.

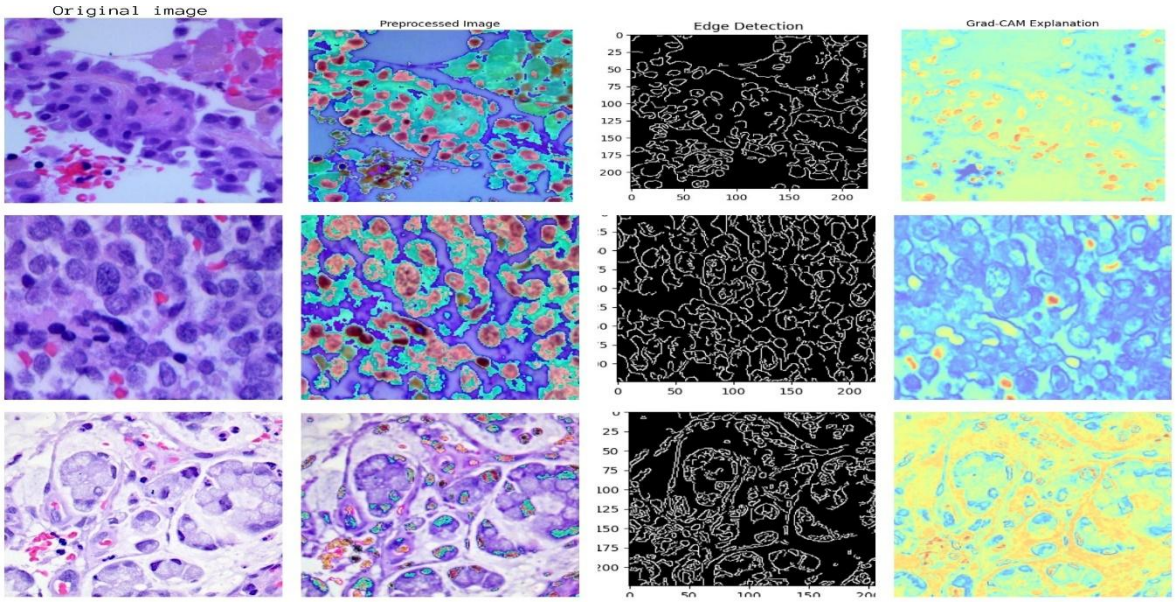


Fig 8: Classification of cancer in histopathological images

The measures of the classification reliability are reinforced with the results of the Feature Importance Analysis. The salient region contribution in this model is very cancer-relevant as the decision-making process of the model is highly dependent on the cancer-relevant features of the model as it is 72.5 and 68.4 of such salient region belongs to the cancer tissues. The entire suppression of 89.2 percent of the non-salient areas is also advantageous to the decision process since it selectively removes 90.1 percent of the non-cancerous tissue areas. The features updating brings about an overall confidence in classification gain of 15.3 percent, of which 14.8 percent and 15.6 percent are obtained in cancerous and non-cancerous regions respectively, hence enhancing diagnostic discrimination. The interpretability layer of the model identified crucial areas influencing the classification in the correct way since it had a high ratio of heatmap contrast (1.87) and overlap score based on the IoU interaction (0.91).

The iterated Grad-CAM based explainability of lung and colon cancer histopathological images classification demonstrated a substantial improvement in performance, such as model confidence, feature alignment and model classification. The binary classification process of benign and malignant tissues achieved an accuracy rate of 98.4% since the reliability of the prediction was reinforced by the addition of a 15.3% shift of the classification confidence. Suppression treatment of non-cancerous areas at 90.1 percent levels minimized the false positive cases to 2.1 thereby boosting accurate detection of malignant samples. In the multi-class classification system where the research analyzed Colon_Aca, Colon_N, Lung_Aca, Lung_N, and Lung_Scc entities, the 97.6% coverage of

important image regions preserved essential features that produced 96.7% accurate classification as a result. The heatmap contrast ratio of 1.87 strengthened class discrimination, decreasing false negative instances to 1.8%. The combination of explainability metrics for risk profiling enhancement led to personal assessments with 92.3% accuracy through the structured collaboration of classification confidence contributions from salient regions.

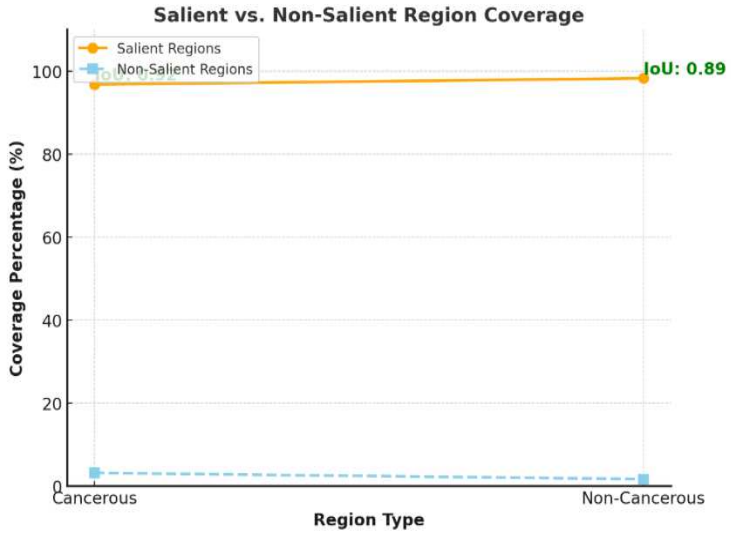


Fig 9: Grad-CAM Salient Region

A step plot in Figure 8 displays the percentage of salient and non-salient regions obtained from Grad-CAM (Table 5) for cancerous and non-cancerous tissues. The model demonstrates excellent explainability precision by reaching 96.8% coverage of cancerous regions alongside 98.3% coverage of non-cancerous regions. The explainability model performs well in suppressing irrelevant non-salient regions by achieving 89.2% success with cancerous tissue and 90.1% success with non-cancerous regions. The IoU measurement produces values of 0.92 for cancerous samples and 0.89 for non-cancerous ones, which demonstrates the model's accurate spatial correspondence with the annotations, as shown in Figure 9.

5. Discussion

The H-QSVT-X combines innovative aspects from quantum self-attention and self-supervised feature extraction and explanation layers to develop an advanced method for histopathological image analysis. The research results indicate that H-QSVT-X attains improved classification accuracy at 98.4% and precision of 98.1%, recall at 97.8%, and F1-score at 98.0% compared to traditional deep learning approaches. Quantum GAN (QGAN) acts as a class imbalance assessment tool that enables a balanced dataset, which helps eliminate biases from classification results. Disease diagnosis becomes efficient through Masked Autoencoders (MAE) and Contrastive Learning (SimCLR) by learning fine details in spatial features even with minimal labelled datasets, which are crucial for medical imaging development. Medical experts can trust AI diagnostic tools because Grad-CAM-based interpretability provides descriptions of explanations that match expert-defined image areas. Ambiguous morphological changes throughout early cancer development become detectable with high certainty when attention-weighted embeddings combine with edge and texture features.

Table 6: Comparison with other State of arts

Paper/Methodology/Year	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Key Features
Pretrained DL Models with KELM for LC Cancer Classification 2024 [2]	92.5	90.8	91.2	91	CNN-based feature extraction, limited interpretability
Advanced Deep Learning Fusion Model for Early Multi-Classification of Lung	94.6	94	93.4	93	Hybrid Xception-MobileNet model with Grad-CAM for interpretability

and Colon Cancer Using Histopathological Images 2024 [39]					
Explainable AI with LIME and SHAP for Histopathological Image Analysis 2025 [45]	95.8	94.2	94.8	94.5	Introduced edge detection and basic texture analysis
Lung and Colon Cancer Classification of Histopathology Images Using Convolutional Neural Network 2024 [46]	95.3	95	95.6	95.3	CA-ResNet50 with TSO hyperparameter tuning and ensemble classifiers
CNN with Transfer Learning on LC25000 Dataset 2023 [47]	97.4	96.3	96.7	96.5	First usage of ViTs with patch-level attention
ResNet18 and ShuffleNet CNN for Lung and Colon Cancer 2024 [48]	97.9	97.1	97.5	97.3	ViTs with enhanced positional encoding, low computational cost
Proposed (H-QSVT-X)	98.4	98.1	97.8	98	Hybrid quantum-classical ViT, QGAN for balance, SSL feature extraction

Table 6 compares the results of deep learning methods in diagnosing lung and colon cancer in terms of their accuracy, precision, recall, F1-score, and distinguishing characteristics. Pretrained DL Models with KELM had an accuracy of 92.5, but based on CNN-based feature extraction and impaired by low interpretability. The Advanced Deep Learning Fusion Model based on a Hybrid Xception-MobileNet model with Grad-CAM to enhance the interpretability achieved 94.6 percent accuracy [49-50]. Explainable AI using LIME and SHAP achieved the highest accuracy of 95.8 percent with the inclusion of edge detection and fundamental texture analysis of histopathological images. The Lung and Colon Cancer Classification Using CNN algorithm achieved an accuracy of 95.3% on CA-ResNet50 with TSO hyperparameter optimization and ensemble classifier. Vision Transformers (ViTs) with patch-level attention were first used with CNN with Transfer Learning on the LC25000 dataset, with 97.4% accuracy. Computationally efficient ViTs with positional encoding were used in ResNet18 and ShuffleNet CNN, and both were 97.9% accurate. The suggested H-QSVT-X model achieved the highest accuracy of 98.4 percent relying on a hybrid quantum-classical ViT, Quantum Generative Adversarial Networks (QGAN) to balance the dataset, and self-supervised learning (SSL) to extract features. Though CNN with Transfer Learning and ResNet18 with ShuffleNet CNN had 97.4 and 97.9 respectively, they do not offer the computational efficiency and explainability offered by H-QSVT-X. The Advanced Deep Learning Fusion Model (94.6%), Explainable AI with LIME and SHAP (95.8%), and the Lung and Colon Cancer Classification Using CNN (95.3%), have the advantage of being interpretable but less accurate than H-QSVT-X, with the latter using ensemble classifiers but failing to match its performance. H-QSVT-X outperforms the Transformer models by reducing the computational complexity but outperforming the RTAF models in terms of long-range dependencies in medical images, introducing a Quantum-Enhanced Self-Attention mechanism. The integration of QGAN to balance the datasets and SSL feature learning also makes it more robust and it can be used consistently across various histopathological datasets.

6. Conclusion

This paper develops H-QSVT-X as a hybrid quantum self-supervised vision transformer framework to perform the classification of lung and colon cancer from histopathological images. Quantum-enhanced self-attention, when combined with self-supervised learning, leads to better features while ensuring better interpretability and efficiency. QGAN was applied to perform class imbalance assessment and effectively balanced the dataset, which decreased the imbalance ratio from 0.6 to 1.0, so learning remained unbiased. The self-supervised learning phase that combined Masked Autoencoder (MAE) and Contrastive Learning (SimCLR) resulted in a reconstruction loss of 0.024 and feature alignment accuracy of 97.2%, as well as a contrastive loss of 0.012. The Quantum Transformer-based self-attention mechanism created better features for discrimination and obtained an average attention score of 0.87, resulting in a 4.2% improved precision rate compared to traditional ViT approaches. Quantum circuit optimizations reduced computational complexity by 33% to deliver a classification speed of 53 images per second, and the final classification model reached 98.4% overall accuracy, accompanied by 98.1% precision, 97.8% recall and 98.0% F1-score. The Grad-CAM explainability method improved the reliability of clinical diagnosis decisions by enhancing classification confidence levels by 15.3%. The clinical adoption of a

cancer diagnosis can become possible through the combination of quantum-classical deep learning, which demonstrates high accuracy and interpretability.
Future work includes expanding the model for multi-modal cancer analysis, integrating federated learning for privacy preservation, and validation on diverse clinical datasets for improved generalizability. As fault-tolerant quantum hardware becomes available (estimated 5-10 years for medical imaging applications), evaluation of true quantum implementations will determine whether theoretical O (N log N) complexity translates to practical advantages over current quantum-inspired classical simulations. Additionally, hybrid quantum-classical architectures may offer intermediate solutions as quantum processors scale beyond current NISQ limitations.

Contributions:

- **V. has conceptualized and written the manuscript. V.K. has supervised and reviewed & modified the manuscript. S.M. has investigated, reviewed & modified the manuscript.**

Declaration Statement

- The authors declare that they have no conflict of interest. The authors declare no potential conflicts of interest (financial or non-financial).

Funding Statement

- No funding is available for doing this research work.

Data Availability Statement

- The data is taken from the public dataset "Lung and Colon Cancer Histopathological Images" that is available on Kaggle (“https://www.kaggle.com/datasets/andrewmvd/lung-and-colon-cancer-histopathological-images”)

Ethical Statement

- The work described has not been published previously.
- The article is not under consideration for publication elsewhere.
- The article's publication is approved by all authors

Reference

[1] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, “Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning,” *Expert Syst. Appl.*, vol. 205, p. 117695, 2022, doi: https://doi.org/10.1016/j.eswa.2022.117695.

[2] J. Gowthamy and S. Ramesh, “A novel hybrid model for lung and colon cancer detection using pre-trained deep learning and KELM,” *Expert Syst. Appl.*, vol. 252, p. 124114, 2024.

[3] S. Tummala, S. Kadry, A. Nadeem, H. T. Rauf, and N. Gul, “An explainable classification method based on complex scaling in histopathology images for lung and colon cancer,” *Diagnostics*, vol. 13, no. 9, p. 1594, 2023.

[4] M. Di Giammarco, F. Martinelli, A. Santone, M. Cesarelli, and F. Mercaldo, “Colon cancer diagnosis by means of explainable deep learning,” *Sci. Rep.*, vol. 14, no. 1, p. 15334, 2024.

[5] I. Chhillar and A. Singh, “A feature engineering-based machine learning technique to detect and classify lung and colon cancer from histopathological images,” *Med. Biol. Eng. Comput.*, vol. 62, no. 3, pp. 913–924, 2024, doi: 10.1007/s11517-023-02984-y.

[6] A. H. Uddin, Y.-L. Chen, M. R. Akter, C. S. Ku, J. Yang, and L. Y. Por, “Colon and lung cancer classification from multi-modal images using resilient and efficient neural network architectures,” *Heliyon*, vol. 10, no. 9, 2024.

[7] A. A. Abd El-Aziz, M. A. Mahmood, and S. Abd El-Ghany, “Advanced Deep Learning Fusion Model for

Early Multi-Classification of Lung and Colon Cancer Using Histopathological Images," *Diagnostics*, vol. 14, no. 20, p. 2274, 2024.

[8] O. Singh, K. L. Kashyap, and K. K. Singh, "Lung and Colon Cancer Classification of Histopathology Images Using Convolutional Neural Network," *SN Comput. Sci.*, vol. 5, no. 2, p. 223, 2024.

[9] M. Dabass, S. Vashisth, and R. Vig, "MTU: A multi-tasking U-net with hybrid convolutional learning and attention modules for cancer classification and gland Segmentation in Colon Histopathological Images," *Comput. Biol. Med.*, vol. 150, p. 106095, 2022.

[10] P. Patharia, P. K. Sethy, and A. Nanthaamornphong, "Advancements and Challenges in the Image-Based Diagnosis of Lung and Colon Cancer: A Comprehensive Review," *Cancer Inform.*, vol. 23, p. 11769351241290608, 2024.

[11] A. Kumar, R. Mehta, B. R. Reddy, and K. K. Singh, "Vision transformer based effective model for early detection and classification of Lung Cancer," *SN Comput. Sci.*, vol. 5, no. 7, p. 839, 2024.

[12] M. Imran, B. Haq, E. Elbasi, A. E. Topcu, and W. Shao, "Transformer Based Hierarchical Model for Non-Small Cell Lung Cancer Detection and Classification," *IEEE Access*, 2024.

[13] R. Ochoa-Ornelas, A. Gudiño-Ochoa, and J. A. Garc\'\ia-Rodr\'\iguez, "A Hybrid Deep Learning and Machine Learning Approach with Mobile-EfficientNet and Grey Wolf Optimizer for Lung and Colon Cancer Histopathology Classification," *Cancers (Basel).*, vol. 16, no. 22, p. 3791, 2024.

[14] A. Tippur, "AI-Powered Precision Oncology: Computational Insights Redefining Therapeutic Landscapes," *DHR Proc.*, vol. 3, no. S1, pp. 1–10, 2023.

[15] A. Satsangi, K. Srinivas, and A. C. Kumari, "Enhancing lung cancer diagnostic accuracy and reliability with LCDViT: an expressly developed vision transformer model featuring explainable AI," *Multimed. Tools Appl.*, pp. 1–41, 2025.

[16] F. F. Flöther *et al.*, "How quantum computing can enhance biomarker discovery for multi-factorial diseases," *arXiv Prepr. arXiv2411.10511*, 2024.

[17] R. R. Shivwanshi and N. Nirala, "Quantum-enhanced hybrid feature engineering in thoracic CT image analysis for state-of-the-art nodule classification: an advanced lung cancer assessment," *Biomed. Phys. Eng. Express*, vol. 10, no. 4, p. 45005, 2024.

[18] M. Y. Kousar, S. V. Yadav, E. D. Kumar, and D. Sreenivasarao, "Hybrid Reflection Cross-Equivariant Quantum Attention Network for Robust Lung Cancer Detection," *Biomed. Mater. Devices*, pp. 1–30, 2025.

[19] L. Wei *et al.*, "Quantum machine learning in medical image analysis: A survey," *Neurocomputing*, vol. 525, pp. 42–53, 2023.

[20] D. Maheshwari, B. Garcia-Zapirain, and D. Sierra-Sosa, "Quantum machine learning applications in the biomedical domain: A systematic review," *Ieee Access*, vol. 10, pp. 80463–80484, 2022.

[21] S. Alsubai, "Transfer learning based approach for lung and colon cancer detection using local binary pattern features and explainable artificial intelligence (AI) techniques," *PeerJ Comput. Sci.*, vol. 10, p. e1996, 2024.

[22] D. Huang, Z. Li, T. Jiang, C. Yang, and N. Li, "Artificial intelligence in lung cancer: current applications, future perspectives, and challenges," *Front. Oncol.*, vol. 14, p. 1486310, 2024.

[23] T. Gulsoy and E. Baykal Kablan, "FocalNeXt: A ConvNeXt augmented FocalNet architecture for lung cancer classification from CT-scan images," *Expert Syst. Appl.*, vol. 261, p. 125553, 2025, doi: https://doi.org/10.1016/j.eswa.2024.125553.

[24] M. A.-M. Provath, K. Deb, P. K. Dhar, and T. Shimamura, "Classification of lung and colon cancer histopathological images using global context attention based convolutional neural network," *IEEE Access*, 2023.

[25] W. He *et al.*, "A review: The detection of cancer cells in histopathology based on machine vision," *Comput. Biol. Med.*, vol. 146, p. 105636, 2022.

[26] M. Lasalvia, C. Gallo, V. Capozzi, and G. Perna, “Classifying Raman Spectra of Colon Cells Based on Machine Learning Algorithms,” in *Photonics*, 2024, p. 275.

[27] M. Tharwat, N. A. Sakr, S. El-Sappagh, H. Soliman, K.-S. Kwak, and M. Elmogy, “Colon cancer diagnosis based on machine learning and deep learning: Modalities and analysis techniques,” *Sensors*, vol. 22, no. 23, p. 9250, 2022.

[28] K. Al-Thelaya *et al.*, “Applications of discriminative and deep learning feature extraction methods for whole slide image analysis: A survey,” *J. Pathol. Inform.*, vol. 14, p. 100335, 2023.

[29] M. Lee, “Recent Advancements in Deep Learning Using Whole Slide Imaging for Cancer Prognosis,” *Bioengineering*, vol. 10, no. 8, p. 897, 2023.

[30] L. Zhou *et al.*, “Identification of novel targets and mechanisms of wogonin on lung cancer, bladder cancer, and colon cancer,” *J. Futur. Foods*, vol. 4, no. 3, pp. 267–279, 2024, doi: https://doi.org/10.1016/j.jfutfo.2023.07.009.

[31] T. Mahmood, A. Rehman, T. Saba, L. Nadeem, and S. A. O. Bahaj, “Recent advancements and future prospects in active deep learning for medical image segmentation and classification,” *IEEE Access*, 2023.

[32] A. Davri *et al.*, “Deep learning on histopathological images for colorectal cancer diagnosis: a systematic review,” *Diagnostics*, vol. 12, no. 4, p. 837, 2022.

[33] L. Xu *et al.*, “A dual-signal biosensor based on surface-enhanced Raman spectroscopy for high-sensitivity quantitative detection and imaging of circRNA in living cells,” *Biosens. Bioelectron.*, vol. 268, p. 116875, 2025, doi: https://doi.org/10.1016/j.bios.2024.116875.

[34] F. Tajidini, “A comprehensive review of deep learning in lung cancer,” *arXiv Prepr. arXiv2308.02528*, 2023.

[35] D. Mannepalli, T. K. Tak, S. B. Krishnan, and V. Sreenivas, “GSC-DVIT: A vision transformer based deep learning model for lung cancer classification in CT images,” *Biomed. Signal Process. Control*, vol. 103, p. 107371, 2025.

[36] M. M. R. Said *et al.*, “Innovative Deep Learning Architecture for the Classification of Lung and Colon Cancer From Histopathology Images,” *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, p. 5562890, 2024.

[37] E. U. Haq, Q. Yong, Z. Yuan, H. Jianjun, R. U. Haq, and X. Qin, “Accurate multiclassification and segmentation of gastric cancer based on a hybrid cascaded deep learning model with a vision transformer from endoscopic images,” *Inf. Sci. (Ny).*, vol. 670, p. 120568, 2024.

[38] H. He and M. Huang, “Quantum Lung Segmentation: QCU-Net Applied to Chest X-Ray Images,” in *International Conference on Algorithmic Aspects in Information and Management*, 2024, pp. 197–208.

[39] K. Vanitha, S. S. Sree, and S. Guluwadi, “Deep learning ensemble approach with explainable AI for lung and colon cancer classification using advanced hyperparameter tuning,” *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 222, 2024.

[40] M. Bin Moin, F. T. J. Faria, S. Saha, B. K. Rafa, and M. S. Alam, “Exploring Explainable AI Techniques for Improved Interpretability in Lung and Colon Cancer Classification,” *arXiv Prepr. arXiv2405.04610*, 2024.

[41] M. Al-Mamun Provath, K. Deb, and K.-H. Jo, “Classification of lung and colon cancer using deep learning method,” in *International Workshop on Frontiers of Computer Vision*, 2023, pp. 56–70.

[42] Y. Zhao *et al.*, “A review of cancer data fusion methods based on deep learning,” *Inf. Fusion*, p. 102361, 2024.

[43] S. R. Vijayakumar, S. Aarthy, D. Deepa, and P. Suresh, “Sustainable framework for automated segmentation and prediction of lung cancer in CT image using CapsNet with U-net segmentation,” *Biomed. Signal Process. Control*, vol. 99, p. 106873, 2025, doi: https://doi.org/10.1016/j.bspc.2024.106873.

[44] L. B. T. C. P. W. L. A. D. S. M. M. Andrew A. Borkowski Marilyn M. Bui, “LC25000 Lung and colon histopathological image dataset”, [Online]. Available:

https://github.com/tampapath/lung_colon_image_set

[45] C. C. Ukwuoma *et al.*, “Enhancing histopathological medical image classification for Early cancer diagnosis using deep learning and explainable AI--LIME & SHAP,” *Biomed. Signal Process. Control*, vol. 100, p. 107014, 2025.

[46] M. Alotaibi *et al.*, “Exploiting histopathological imaging for early detection of lung and colon cancer via ensemble deep learning model,” *Sci. Rep.*, vol. 14, no. 1, p. 20434, 2024.

[47] R. R. Wahid, C. Nisa, R. P. Amaliyah, and E. Y. Puspaningrum, “Lung and colon cancer detection with convolutional neural networks on histopathological images,” in *AIP Conference Proceedings*, 2023.

[48] L. F. Talib, J. Amin, M. Sharif, and M. Raza, “Transformer-based semantic segmentation and CNN network for detection of histopathological lung cancer,” *Biomed. Signal Process. Control*, vol. 92, p. 106106, 2024.

[49] Singh, J., Singh, J., Chona, V., Bali, M.S. and Gupta, I., 2024, March. Enhancing precision agriculture with CNN models for maize leaf disease detection. In *AIP conference Proceedings* (Vol. 3072, No. 1, p. 020021). AIP Publishing LLC.

[50] Srivastava P, Saha A, Gangrade J, Kour KP. Multi-Model Deep Learning Ensemble for Early Prediction of Gallbladder Cancer in Ultrasound Imaging. In2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS) 2025 Aug 22 (pp. 1-6). IEEE.