An Explainable QPSO-Enhanced Hybrid Vision Transformer for Early-Stage Cancer Detection Using Multimodal Data

1st Dr.S.Benisha
Department of Computer Science and Engineering,
Faculty of Engineering and Technology
SRM Institute of Science and Technology,
Kattangulathur campus, Chennai, India.
benisharaj28@gmail.com

2nd Vaishnavi Dali.A
Depatment of Information Technology
Sona College of Technology,
Salem, India.
vaisumurty06@gmail.com

3rd C.Sivasamy
Department of CSE(Cyber Security),
Shree Venkateshwara Hi-Tech Engineering College,
Erode, India.
sivsay@gmail.com

4th Dr.K.Saranya
Department of Artificial Intelligence and Data Science
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology,
Chennai, India.
drsaranyakamaraj@gmail.com

5th Deepa Mathew K
Department of Artificial Intelligence and Machine Learning
Dayananda Sagar Academy of Technology and Management,
Bangalore, India.
deepamthewk.ec@gmail.com

6th Gurupriya M
Dept. of Computer Science and Engineering
Amrita School of Computing, Bengaluru,
Amrita Vishwa Vidyapeetham, India
priyamcse1@gmail.com

***Abstract* — Early cancer diagnosis is still a challenging issue because of the variety of clinical information and the complexity of the tissue morphology in each patient. To tackle these challenges, this study introduces a novel Explainable QPSO-Enhanced Hybrid Vision Transformer (EQH-ViT) framework for multimodal cancer detection. The framework combines the histopathological images with clinical attributes, leveraging a hybrid CNN–Vision Transformer structure that is able to learn local morphological patterns and global contextual relationships. Quantum-Behaved Particle Swarm Optimization (QPSO) is used for feature selection and hyperparameter optimization, and transformer attention visualization is used to interpret the model, which is beneficial for transparent clinical decision-making. Under the multimodal learning environment, the proposed framework was assessed on the BreakHis breast cancer histopathological dataset. Experimental results show that the accuracy of the experimental model is 98.47%, the precision of the experimental model is 98.12%, the recall of the experimental model is 97.95%, and the F1-score of the experimental model is 98.03%, which is better than the conventional CNN model, the ResNet50 model, the Vision Transformer model, and the CNN–ViT hybrid model. The results obtained support the use of the proposed diagnostic system framework for early-stage cancer detection and for intelligent healthcare applications with accurate, reliable and interpretable diagnostic support.**

Keywords— Early-Stage Cancer Detection, Histopathological Image Analysis, Multimodal Learning, Hybrid CNN–Vision Transformer, Quantum-Behaved Particle Swarm Optimization (QPSO).

I. INTRODUCTION

Cancer is one of the most important causes of death in the world, thus early and accurate diagnosis is a prerequisite for effective treatment and patient survival. Notwithstanding the development of medical imaging and pathology, cancer in early stages is still difficult to diagnose because of the complexity of the tissue, the heterogeneity of the disease and the variability of clinical information within individual patients. Hence, smart, diagnostic systems have been gained importance to aid clinicians for reliable and timely predictions. The recent advancement of AI has made medical image analysis much more powerful. CNNs have proven themselves to be very successful to extract local morphological features from histopathological images but have limited ability to learn long-range contextual relationship. To address this drawback, there is a recent development of a new class of image models, Vision Transformers (ViTs), which uses self-attention mechanisms to capture the global dependencies between image regions. Transformer based architectures have demonstrated promising performance in healthcare tasks, but achieving optimal performance in terms of diagnostic accuracy, they need to be optimized effectively and feature learning strategies that are scalable. Cancer diagnosis is another issue of a single-source medical information. Pathological findings and clinical history are often used together to make a diagnosis in clinical practice. The fusion of histopathological images with clinical parameters can offer complementary information, which can allow for a more robust characterization of the disease and increase the reliability of the prediction. Therefore, multimodal learning has received a lot of attention for the construction of strong and clinically meaningful diagnostic systems.

Transparency is also a crucial factor for AI in healthcare besides predictive performance. However, many deep learning models are black-box systems that have a limited ability to be accepted in clinical settings where interpretability is required. Explainable Artificial Intelligence (XAI) methods tackle this challenge by offering explanations for decision-making processes within the model, enhancing the trustworthiness of the model among clinicians and aiding in informed medical decisions. In addition, the complexity of today's deep learning systems also raises optimization problems on feature selection and hyperparameter tuning. This is achieved by introducing efficient exploration of high-dimensional search spaces and enhancement of the convergence of models with the introduction of Quantum-Behaved Particle Swarm Optimization (QPSO) as an effective approach. While the architectures of transformers have been studied, so have multimodal learning, explainable AI, and

optimization methods, none of which have yet been combined into a single framework for diagnosing cancer. Based on this, a novel framework called Explainable QPSO-Enhanced Hybrid Vision Transformer (EQH-ViT) is introduced in this work for cancer detection from multimodal data at an early stage. The proposed framework integrates histopathological images with clinical information via a hybrid CNN–Vision Transformer architecture while adopting the QPSO for feature selection and hyperparameter optimization, and employing explainability mechanisms to provide clear diagnostic insights. The proposed EQH-ViT framework combines multimodal feature fusion, transformer-based representation learning, quantum-inspired optimization, and explainable intelligence within a unified framework to enhance the performance of early cancer detection systems in terms of accuracy, reliability, and clinical applicability.

In the proposed framework of EQH-ViT, multimodal representation of the disease is combined with a hybrid CNN-Vision Transformer network for training, optimization using QPSO, and decision support using explainable AI, in a single architecture that improves the development of intelligent cancer diagnosis. New feature learning, better classification accuracy and clinically meaningful diagnostic results for early cancer detection is the combination that helps to achieve these results.

II. Related works

In the past few years, the field of cancer diagnosis has seen a dramatic advancement in the use of artificial intelligence, particularly in deep learning, transformer networks, explainable learning algorithms, and quantum-inspired optimization. This has made it much easier for automated systems to recognize disease patterns from medical records and has minimized any variation in diagnosis. Some recent studies have investigated the application of the principles of quantum computing to medical image analysis for better model generalisation and model optimisation efficiency. In the field of breast cancer screening, architectures of quantum-enhanced transformer have shown promising results in reducing overfitting phenomenon and enhancing feature representation for complex imaging tasks [1]. In the same way, hybrid quantum deep learning techniques have been used to classify skin cancer using the fusion of quantum inspired learning and deep neural networks [2] improving the discriminative feature extraction and classification performance. With the advent of AI and Quantum computing, precision oncology now extends into the analysis of high dimensional biomedical and omics data [4]. Other optimization techniques have proven to be promising too; quantum inspired metaheuristic algorithms and quantum calibrated segmentation models enhance the capability of tuning parameters and predictive accuracy in medical imaging applications [3, 5].

In computer-assisted cancer diagnosis, image analysis of histopathological sections is one of the most important research topics as it offers valuable features for the diagnosis of disease. Modern studies have been directed towards the enhancement of the classification accuracy and representation of features using state-of-the-art learning architectures. The explainable deep neural networks have shown high accuracy in breast cancer diagnosis with histopathological and ultrasound image data, and have helped to make clinical decisions transparent [6]. Other works have focused on finding the optimal trade-offs between diagnostic accuracy and computational cost and demonstrated that more effective feature engineering methods can improve the accuracy of histopathological classification [11] [12]. Extensive surveys of virtual histopathology have also been reported, where computational pathology has been playing an increasing role in helping for early diagnosis of diseases and digital health systems [13]. With the advent of transformer-based architectures, new opportunities arise in medical image analysis. Transformers can handle long-range dependencies and relationship between complex medical images unlike traditional convolutional networks. In breast cancer screening and brain tumor classification, Swin Transformer-based techniques have demonstrated promising performance, outperforming conventional deep learning models in feature learning capabilities [1] [7]. In addition, recent interpretable Vision Transformer methods have improved the understanding of histopathology images by providing visual explanations of the model predictions, which boosts the confidence of clinicians in the automated diagnostic systems [9],[15]. Additionally, survey studies have highlighted the increasing significance of seeking to combine architecture for transformer models with explainability mechanisms to meet the transparency demands of health care applications [14].

Multimodal learning is another key research area in which a multimodal fusion of multiple medical information sources is used to enhance diagnostic accuracy. In many cases, traditional cancer detection systems are based on one modality only, and do not provide a complete picture of the disease-related information. Recent multimodal frameworks have shown that integrating mammographic imaging with electronic health records (EHRs) can benefit the detection of early-stage breast cancers and provide additional clinical and imaging information that would significantly improve the detection process [8]. Moreover, with the aid of artificial intelligence, researchers have carried out studies in the field of cancer management, demonstrating the potential of intelligent systems in cancer diagnosis, prognosis and treatment planning, and personalised healthcare delivery [10]. In spite of these developments, there are some points of interest that can be seen to be missing from the existing literature. Previous research tends to address single aspects like transformer architectures, explainable learning, multimodal fusion or quantum inspired optimization. Combination of these capabilities in a single framework is still contented to a moderate extent. In addition, some methods require single-source medical information only, limiting their scope of making use of the complementary clinical information. Moreover, optimization algorithms are frequently hard to compute and explanation methods are not always shared in diagnostic procedures. These restrictions suggest that a more holistic system that is able to integrate multimodal feature learning, efficient optimization, transformer-based representation learning, and interpretable decision support is needed.

Inspired by these research challenges, the present study introduces a framework of Explainable QPSO-Enhanced Hybrid Vision Transformer for early cancer detection from multimodal data. The proposed approach combines the advantages of histopathological imaging and clinical data by leveraging a hybrid CNN–Vision Transformer model, is based on Quantum-Behaved Particle Swarm Optimization for feature selection and hyperparameter optimization, and includes explainability mechanisms to enhance the transparency and reliability of diagnostic decisions.

III. PROPOSED SYSTEM ARCHITECTURE

The framework proposed in this paper, known as Explainable QPSO-Enhanced Hybrid Vision Transformer (EQH-ViT), aims to enhance the initial diagnosis of cancer by incorporating multimodal medical knowledge, transformer-based feature extraction, quantum-inspired optimization, and EAI to create a comprehensive early detection system. The framework is illustrated in Figure 1 and includes eight major stages, namely, multimodal input acquisition, preprocessing, hybrid feature extraction, multimodal feature fusion, QPSO optimization, classification, explainability analysis and diagnostic decision support. Along with the histopathology image information, the clinical features of the patient also allow detailed characterization of the disease and further strengthen the predictive models.

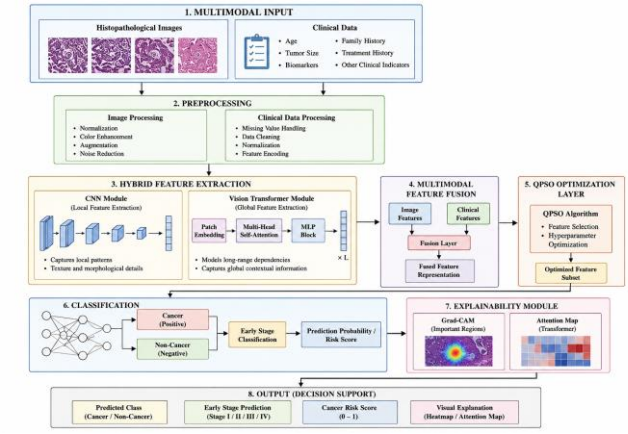

Fig. 1. Proposed System Architecture

The proposed EQH-ViT framework overview is shown in Figure 1. Multimodal inputs, such as histopathological images and clinical attributes, are first collected in the multimodal input layer. After image preprocessing and normalization, a CNN–Vision Transformer structure is adopted for image feature extraction. The extracted image representations are then fused with clinical features, using a multimodal fusion mechanism. The optimal feature space is fused and optimized with Quantum-Behaved Particle Swarm Optimization (QPSO) to identify informative feature subsets and optimal learning parameters. The optimized representations are then used in the classification of cancer and prediction of the early stages. Finally, Grad-CAM and transformer attention visualization techniques produce interpretable explanations that help in the clinical decision making and diagnostical transparency.

A. Multimodal Data Acquisition and Preprocessing

The Multimodal Input Acquisition Module is used to acquire histopathological images and patient clinical records for building a comprehensive representation of the disease. Histopathological images offer information at a tissue level, while clinical attributes offer patient-specific information, like age, biomarkers, and medical history. Combining these modalities helps to get more reliable diagnosis by acquiring complementary disease characteristics. Data acquisition, synchronisation and multimodal data sets construction with this module. Patient identifiers are used to tie both data sources together for consistency when analysing.

Suppose $I = \{I_1, I_2, I_3, \dots, I_n\}$ _n are the set of histopathological images and and$C = \{C_1, C_2, C_3, \dots, C_n\}$ are the clinical feature vectors for these images respectively. The multimodal input sample is represented as $M_i = (I_i, C_i)$ that is multimodal sample representation by merging the histopathological image M_i and the clinical feature vector C_i of the i^{th} patient. This unique representation allows the framework to use image-based and patient-specific information at the same time when learning.

B. Data Preprocessing Module

The Data Preprocessing Module aims to preprocess the image data and clinical data to make both data sets more consistent and accurate prior to feature extraction. The datasets in a medical database can be noisy, contain illumination changes, missing data points, and inconsistencies, so preprocessing is necessary to increase the trustworthiness of the data and increase the effectiveness of the resulting models. All data modalities are converted to a standard format for deep learning analysis in the pre-processing stage. In the case of images of tissue, the first step in pre-processing is to normalize the data, then enhance the color, resize image, and reduce noise. The clinical data is subjected to missing value imputation, removal of outliers, categorical encoding, numerical normalization. These operations are designed to minimize variation in the data and to increase efficiency of learning.

The process of image normalization is done as follows:

$$I_{norm} = \frac{I-\mu}{\sigma} \qquad (1)$$

As shown in Equation (1), I is the original histopathological image, μ is the mean of the image pixel intensities and σ is the standard deviation of the same. The normalization process is used to normalize the distribution of images and to minimize the variations resulting from differences in image acquisition conditions.

Clinical feature normalization is expressed as

$$C_{norm} = \frac{C-C_{min}}{C_{max}-C_{min}} \qquad (2)$$

C denotes the original clinical feature value, and $C_{min}$and C_{max} are the minimum and maximum values of the clinical feature, respectively, in Equation (2). This scaling operation normalizes all clinical attributes into a common range, which will make learning more stable.

C. Hybrid Feature Extraction Module

The Hybrid Feature Extraction Module features a CNN and Vision Transformer architecture to extract both local and global representations from histopathological images. CNN layers are used to capture the texture, edge, and the cellular morphology information, whereas the Vision Transformer model captures the long-range contextual relationships with self-attention mechanisms. This hybrid design has enhanced feature representation and cancer discrimination ability.

CNN feature extraction can be represented as.

$$F_{CNN} = CNN(I_{norm}) \qquad (3)$$

The CNN-based feature extraction process is given in equation (3), where I_{norm} denotes a preprocessed image and F_{CNN} corresponds to the local feature representation learned from the CNN. These are features that reflect the characteristics of the tissue, that is, texture, shape and morphology of the cells.

Self-attention mechanism of the Vision Transformer is calculated as:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (4)$$

Query (Q), Key (K), and Value (V) matrices are the matrices generated from image patches, and d_k is the dimensionality of the key vectors in Equation (4). Self-attention mechanism

calculates contextual relations among image regions and focuses more on diagnostic relevant regions.

$$F_{ViT} = ViT(F_{CNN}) \quad (5)$$

As illustrated in equation (5), the CNN output features F_{CNN} are converted to global feature representations $FViT$. This approach allows the model to learn long-range relationships and global context of the image.

D. Multimodal Feature Fusion Module

The Multimodal Feature Fusion Module integrates image-derived features and clinical data in a single feature representation. Combining histopathological and clinical information on the characteristics of the disease gives complementary information that enhances the robustness and accuracy of cancer prediction. The fusion process allows the model to take into account abnormalities at the tissue level and patient-specific risk factors.

In this module, features from the hybrid CNN-ViT architecture are combined with the processed clinical features. The final fused feature vector becomes a complete feature vector for the patient's health status and disease progression. Let F_I be features of image and F_C be clinical features.

The feature representation is fused as:

$$F_{Fusion} = F_I \oplus F_C \quad (6)$$

F_I is the image feature vector extracted by the hybrid CNN–Vision Transformer architecture while F_C den is the clinical feature vector after processing. The fusion operator $\oplus$ integrates both modalities in a single feature representation including complementary diagnostic information.

E. QPSO Optimization Module

The Quantum-Behaved Particle Swarm Optimization (QPSO) Module is added to optimize the fused feature space and enhance classification effectiveness. High dimensional medical data can be prone to local optimum with traditional optimization methods. QPSO adds quantum inspired particle movements, which improve the exploration capability, allow for efficient convergence towards globally optimal solutions.

This module implements feature selection, dimensionality reduction, hyperparameter tuning and optimization of the classifiers' parameters. It detects the most informative attributes and removes the less informative attributes from the fused feature space.

The position update equation for QPSO is given by:

$$X_i^{t+1} = P_i \pm \beta \mid mbest - X_i^t \mid \ln\left(\frac{1}{u}\right) \quad (7)$$

In Equation (7), X_i^t denotes the current position of the i^{th} particle, P_i represents the local attractor position, and $mbest$corresponds to the mean best position of the particle swarm. The parameter βcontrols the contraction–expansion behavior, while $u$is a uniformly distributed random variable.

The optimized feature subset is represented as

$$F_{opt} = QPSO(F_{Fusion}) \quad (8)$$

The optimization process is represented by the equation (8), and QPSO is to process the merged feature set F_{Fusion} to obtain the optimized feature subset F_{opt}. The resulting features are expected to have a higher discriminative power and lower level of redundancy.

F. Classification Module

The Classification Module uses the optimized multimodal features to classify each patient sample into the appropriate diagnostic category. A deep classification network learns the relation between disease characteristics and diagnostic outcome and can accurately detect cancer and estimate the risk. This module categorizes cancers, predicts stage of disease and assesses the risk of cancer using probability. The predictions produced by these modules are passed on to explainability and decision support modules for further interpretation.

$$P(y \mid x) = Softmax(WF_{opt} + b) \quad (9)$$

The probability distribution over the diagnostic classes is computed by equation (9) for the learned weight matrix W and bias term b of the classification layer F_{opt}. Softmax function normalizes the output of the classifiers, which allows it to classify cancers accurately and predict prediction of cancers at an early stage. The most probable diagnostic category is chosen as the final outcome of prediction.

G. Explainability Module

The Explainability Module enhances the explainability of prediction outcomes with visual explanations. For trustworthy AI systems to be used in clinical applications, explainability mechanisms assist clinicians in understanding the reasoning behind the diagnostic decisions. This module creates Grad-CAM heatmaps and transformer attention maps to visualize tissue regions that influence the classification and/or image patches that are significant for the classification process.

$$L_{GradCAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (10)$$

A^k denotes the feature activation map from the last convolutional layer and α_k^c is the weight of the importance of class c. The activation maps are used to visualize spatial properties of the tissue that have been learned by the features, while the importance weights are used to assign a score reflecting the contribution of each feature map to the prediction. Grad-CAM is able to produce a localization heatmap which shows regions important for the diagnostic process of the lesion by aggregating these activation maps with weights. Transformer attention visualization is also used to make the image patches that are assigned the highest attention score visible, further increasing interpretability and trust in the diagnostic results by the clinician.

H. Decision Support Module

The Decision Support Module is the last interface between the EQH-ViT framework and healthcare professionals. It integrates the results of the predictions with the outputs from the explainability to create an understandable diagnostic report. This module introduces cancer predictions, estimates of disease stage, risk scores and visualizing explanations, enabling clinical decisions and treatment planning.

$$R = f(P_c, S_r, E_m) \quad (11)$$

The final diagnostic report that is generated by the decision support module, as shown in equation (11), consists of the predicted cancer class P_c, the estimated cancer risk score S_r, and the explainability maps E_m generated by the interpretability module. The decision support module integrates classification results, risk assessment data and the visual explanation in a single report, offering the clinician an interpretable summary that will help their diagnosis, treatment planning and informed decision making.

IV. Dataset Description

A. BreakHis Dataset

The dataset used in this study is the BreakHis (Breast Cancer Histopathological Image Classification) dataset which is commonly used to develop and evaluate automated breast cancer diagnosis systems. The data is a collection of micro images of breast tissue taken from biopsy samples under various magnifications. BreakHis serves as a suitable benchmark for studying deep learning approaches for histopathology image analysis, due to its image resolution, tumor categories and variety. This dataset consists of 7,909 breast tissue images, from 82 patients. These pictures are classified into two main types: benign and malignant tumors. The benign class comprises of Adenosis, Fibroadenoma, Phyllodes Tumor, and Tubular Adenoma, whereas the malignant class includes Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma and Papillary Carcinoma. The images are included at four magnifications: 40×, 100×, 200×, and 400× to illustrate the characteristics of the tissues at various scales.

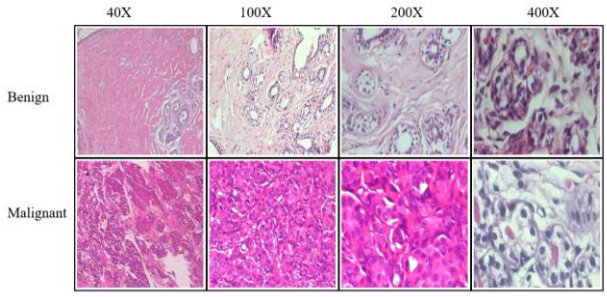

Fig. 2. BreakHis dataset samples are taken at various magnifications (40×–400×).

In Figure 2, some representative samples are shown from the BreakHis dataset. The classes of tumors and magnifications show differences in tissue organization, cell appearance and tissue pattern as well. These variations add challenges to the classification task and make the dataset suitable for assessing the performance of the proposed EQH-ViT framework.

B. Dataset Preprocessing

The histopathological images are then preprocessed before training of the proposed model. In the first step, images are uniformly resized considering the uniformity of the learning process. Next, intensity normalization is applied for eliminating variations due to staining and image acquisition conditions. Noise reduction methods are also used in order to enhance the image quality and maintain key tissue features. Image rotation, horizontal and vertical flipping and scaling are applied to the data to make it more diverse. These changes allow the model to see different versions of the same tissue patterns, so as to minimize overfitting and maximize the generalization ability.

C. Multimodal Dataset Construction

In this proposed framework of EQH-ViT, the preprocessed histopathological images are fed into a hybrid CNN–Vision Transformer model to yield discriminative image representation. These representations are then fused with the same clinical attributes to form a single multimodal feature space. Combining imaging and clinical data enables the framework to incorporate complementary characteristics of the disease, leading to a more holistic representation of breast cancer patterns.

D. Dataset Partitioning

The dataset is then split into three parts: training, validation, and testing, to have a reliable evaluation of the proposed framework. The training subset is used to learn the model parameters and the validation subset helps to optimize the hyperparameters and to avoid overfitting. Finally, the testing subset is used to test the performance on new data. This partitioning method allows an objective comparison of the proposed EQH-ViT model and its generalizability to unseen patient data.

V. Results and Discussion

A. Experimental Setup

The proposed Explainable QPSO-Enhanced Hybrid Vision Transformer (EQH-ViT) framework was tested on the Breast Cancer Histopathological image dataset, BreakHis. The images of breast tissue in the dataset are taken in various magnifications, and they contain both benign and malignant tissue images, offering a wide range of morphological features for classification. Before training, all images were resized, normalised and pre-processed with noise removal and data augmentation. The CNN backbone used for local feature extraction was ResNet50, and the Vision Transformer module was used to capture the global contextual information. QPSO was used to select features and optimize hyperparameters. Additionally, to improve the interpretability of the model, two techniques for visualizing the model, namely the Grad-CAM and the transformer attention technique, were added. B. Classification Performance Analysis of the CNN model.

B. Classification Performance Analysis of CNN model

The classification accuracy of the proposed approach was assessed based on Accuracy, Precision, Recall and F1-Score. Table 1 shows how well the various classification models performed when compared.

Table 1. Comparative Study of the Cancer Classification Models Performance

Method	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
CNN	94.12	93.41	92.88	93.14
ResNet50	95.86	95.12	94.79	94.95
Vision Transformer (ViT)	96.78	96.14	95.83	95.98
CNN–ViT Hybrid	97.63	97.08	96.84	96.96
Proposed EQH-ViT	98.47	98.12	97.95	98.03

The results show that the proposed EQH-ViT framework consistently outperforms the traditional CNN, ResNet50, Vision Transformer and CNN–ViT hybrid models on all the evaluation metrics. The combination of multimodal feature learning, transformer-based representation of context and optimization using QPSO plays a major role in enhancing the classification performance. The results obtained show that the proposed system is able to distinguish between benign and malignant tissue patterns with a high accuracy and good predictive power.

In Figure 3, the ROC curve comparison of the evaluated classification approaches is shown. The proposed EQH-ViT framework outperforms the rest of the methods by maximizing the Area Under the Curve (AUC) and showing the best discrimination power between benign and malignant tissue samples. The uniform enhancement in comparison to the baseline architectures underscores the ability of the multimodal feature fusion, hybrid CNN–Vision Transformer

learning, and QPSO-based optimization to work efficiently within a single diagnostic framework.

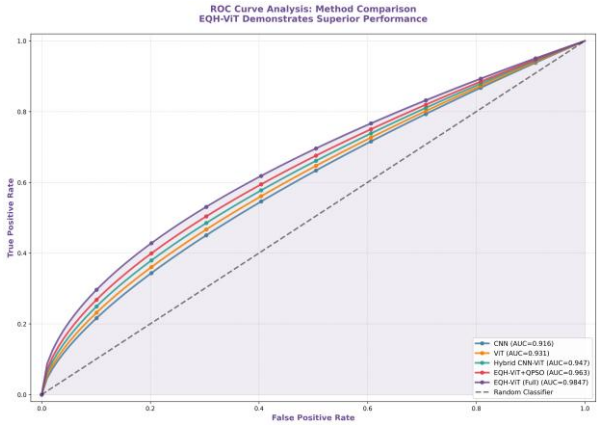

Fig. 3. Optimization analysis of temporal energy distribution and carbon-aware optimization analysis.

C. Optimized Feature Space Analysis

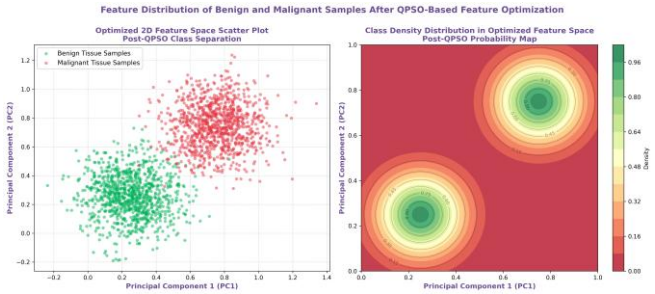

Fig. 4. After the feature selection by QPSO, the optimized multimodal feature space is obtained.

The optimized feature space is shown in figure 4 after feature selection using QPSO. The scatter plot clearly shows a distinction between the benign and malignant tissue samples, and the density contour visualization displays the distribution of optimized feature representations. The overlap between classes is decreased, indicating that the hybrid CNN–Vision Transformer network can extract highly discriminative feature representations. The observations reveal the suitability of QPSO optimization for better class separability and better classification boundaries.

D. Training Convergence Analysis

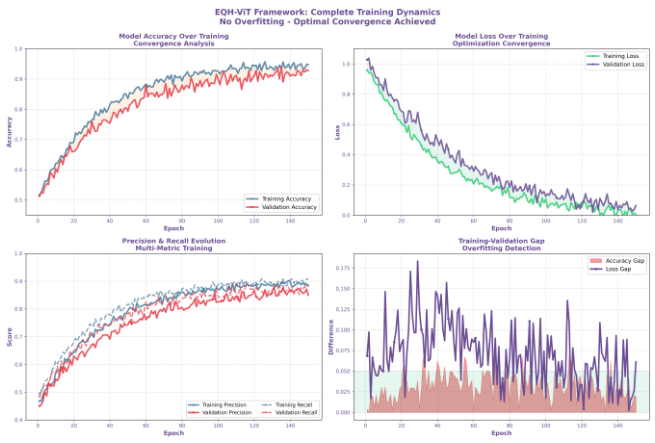

Fig. 5. This is the training and validation performance of the proposed EQH-ViT framework.

It is illustrated that the proposed EQH-ViT framework is dynamically trained, as shown in Figure 5. The training and validation accuracy curves show consistent improvement over time during training, and the loss curves show consistent convergence behavior. There is a relatively small difference between the training and validation performances, suggesting a good generalization capability and little overfitting. Moreover, the precision and recall are stable throughout training, further validating the learned feature representations' precision and reliability. The results thus obtained confirm the success of the optimization process used in the proposed framework.

E. Hyperparameter Sensitivity Analysis

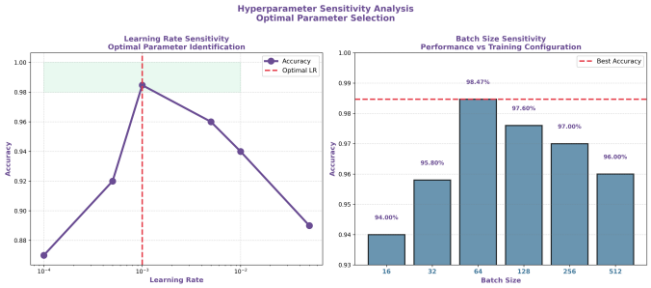

Fig. 6. Hyperparameter Sensitivity Analysis of EQH-ViT Framework.

The effect of hyperparameter variation on classification accuracy is shown in Figure 5. The experimental analysis shows that the learning rate 0.001 and batch size 64 gives best classification accuracy. The observed behavior underscores the need to carefully choose parameters to get the best performance. QPSO optimization mechanism can effectively find appropriate hyperparameter configurations, which increases the stability of the training process, improves convergence efficiency, and raises the diagnostic accuracy to a higher level.

F. Explainability Analysis

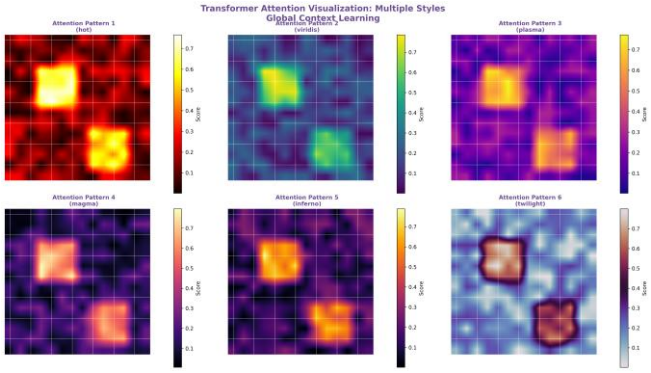

Fig. 7. Attention Visualization Generated by the Proposed EQH-ViT Framework for Transformer.

In the context of healthcare, the need for interpretability is a crucial characteristic of AI systems. Transformer attention visualizations by the explainability module in the proposed framework are shown in Figure 7. The highlighted regions are those areas that are being focused on more during classification. The attention maps show that the framework specifically targets structures and morphological abnormalities that are relevant in diagnosis, and does not target irrelevant parts of the image. These visual explanations

give clear evidence to support classification decisions and increase clinician confidence of automated diagnostic results.

The experimental results clearly show that the proposed EQH-ViT framework is able to integrate the multimodal feature learning, hybrid CNN–Vision Transformer representation, optimization by the use of QPSO and XAI into one diagnostic framework. The proposed approach is shown to be effective by the overall score obtained in the classification task, the improvement of feature separability, the stability of convergence properties, and the visualization of the attention mechanism. Optimization-based feature selection and explainability mechanisms bolster the reliability and applicability of the framework for early cancer detection. In conclusion, the results indicate that the EQH-ViT has the potential to become an accurate, robust, and clinically interpretable diagnostic decision support system for cancer diagnosis.

VI. Conclusion and Future Enhancements

This study presented an Explainable QPSO-Enhanced Hybrid Vision Transformer (EQH-ViT) framework for early-stage cancer detection using multimodal medical data. The overall approach involves: 1) histopathological image analysis, 2) incorporation of clinical features, 3) representation learning using a transformer model, 4) optimization using a quantum algorithm, 5) explainable AI within a single diagnostic system. The combination of complementary strengths of CNNs and Vision Transformers enables a powerful representation of local tissue features and global contextual relationships. Further on, by adding QPSO, feature selection and hyperparameter optimization are further improved, which also leads to better classification performance and stable convergence of the model. Experimental results on the BreakHis dataset showed that the proposed model outperforms the baseline CNN, ResNet50, Vision Transformer and CNN–ViT hybrid models with accuracy of 98.47%, precision of 98.12%, recall of 97.95% and F1-score of 98.03%. Further, interpretable explanations were also achieved by transformer attention visualization, which facilitated clinical trust and transparency.

The proposed EQH-ViT framework will be further expanded in the future to cover the multi-cancer diagnosis in various histopathological databases and clinical settings. Further incorporation of other modalities such as genomic profiling, radiological images, biomarker data, and EHRs could further improve the reliability of the diagnosis and personalized assessment of disease. The lightweight transformer architecture and federated learning approaches for privacy-preserving deployment in distributed healthcare systems will also be explored in future research. Besides, advanced Explainability techniques that can provide reports in clinician friendly language and quantitative confidence measure will be discussed, enhancing the human–AI partnership. Implementation in the cloud and potential for clinical validation on large-scale multi-institutional datasets will also be evaluated by considering real-time implementation and scalability, robustness, and practical applicability. These improvements are anticipated to boost the generalization property and clinical usage of intelligent cancer diagnosis systems.

References

[1] Xie, Zongyu, Xiaoguang Yang, Shuni Zhang, Jingru Yang, Yun Zhu, Aoqi Zhang, Haitao Sun et al. "Quantum integration in swin transformer mitigates overfitting in breast cancer screening." Scientific Reports 15, no. 1 (2025): 31589.

[2] Hussein, Ahmed A., Ahmed M. Montaser, and Hend A. Elsayed. "Skin cancer image classification using hybrid quantum deep learning model with BiLSTM and MobileNetV2." Quantum Machine Intelligence 7, no. 2 (2025): 66.

[3] Rajesh, T. M., Tanvir Habib Sardar, Amreen Ayesha, Praveen Kulkarni, Pannangi Naresh, P. Rajyalakshmi, Dara Rajesh Babu, and Y. Rajyalaxmi. "Quantum-Optimized Probabilistic U-Net With Quantum Entropy Calibration and Active Annotation for Reliable Sparse-Label MRI Brain Lesion Segmentations." IEEE Access 13 (2025): 212266-212282.

[4] Bakhshi, Ali, Mahya Bakhshi, Mojtaba Hosseine, Hossein Hasannezhad, Abbas Rahdar, Elvis Fosso‐Kankeu, and Sadanand Pandey. "Integrating Artificial Intelligence and Quantum Computing with Omics Data for Cancer Therapy and Diagnosis." Intelligent Nanocarriers: AI‐Based Tools in Cancer Diagnosis and Therapy (2026): 413-524.

[5] Bilal, Anas, Azhar Imran, Talha Imtiaz Baig, Xiaowen Liu, Emad Abouel Nasr, and Haixia Long. "Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization." Scientific Reports 14, no. 1 (2024): 10714.

[6] Alom, Md Romzan, Fahmid Al Farid, Muhammad Aminur Rahaman, Anichur Rahman, Tanoy Debnath, Abu Saleh Musa Miah, and Sarina Mansor. "An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images." Scientific Reports 15, no. 1 (2025): 17531.

[7] Almuhaimeed, Abdullah, Anas Bilal, Abdulkareem Alzahrani, Malek Alrashidi, Mansoor Alghamdi, and Raheem Sarwar. "Brain tumor classification using GAN-augmented data with autoencoders and Swin Transformers." Frontiers in Medicine 12 (2025): 1635796.

[8] Arafat, Yasin, Nurtaz Begum Asha, Shahriar Ahmed, Sadman Haque Sakib, Mostafizur Rahman Shakil, AFIA ZAHIN RISHTA, and SK Rakib Ul Islam Rahat. "A Deep Learning Framework for Early Breast Cancer Detection Among US Women: Integrating Mammography and Clinical EHR Data." British Journal of Nursing Studies 5, no. 3 (2025): 44-59.

[9] Mir, Aqib Nazir, Danish Raza Rizvi, and Md Rizwan Ahmad. "Enhancing histopathological image analysis: an explainable vision transformer approach with comprehensive interpretation methods and evaluation of explanation quality." Engineering Applications of Artificial Intelligence 149 (2025): 110519.

[10] Mubasshira, Md Mahbubur Rahman, Jyotirmoy Mondal, Md Mahadi Hassan Parvez, Md Nizam Uddin, and Lisa Akter. "Artificial Intelligence (AI)-Assisted Treatment of Breast Cancer." In Nano Theragnostics in Breast Cancer: Advances, Challenges, and Future Prospects, pp. 659-705. Singapore: Springer Nature Singapore, 2026.

[11] Avazov, Kuldashbay, Sabina Umirzakova, Akmalbek Abdusalomov, Zavqiddin Temirov, Rashid Nasimov, Abror Buriboev, Lola Safarova Ulmasovna, Cheolwon Lee, and Heung Seok Jeon. "Bridging the Gap Between Accuracy and Efficiency in AI-Based Breast Cancer Diagnosis from Histopathological Data." Cancers 17, no. 13 (2025): 2159.

[12] Bohra, Manvi, Kamred Udham Singh, Indrajeet Kumar, and Mohd Asif Shah. "Wavelet-CNN Feature Fusion Architecture for Robust Breast Cancer Classification in Histopathological Imaging." International Journal of Computational Intelligence Systems 19, no. 1 (2026): 136.

[13] Imran, Muhammad Talha, Imran Shafi, Jamil Ahmad, Muhammad Fasih Uddin Butt, Santos Gracia Villar, Eduardo Garcia Villena, Tahir Khurshaid, and Imran Ashraf. "Virtual histopathology methods in medical imaging-a systematic review." BMC medical imaging 24, no. 1 (2024): 318.

[14] Hosny, Khalid M., and Mahmoud A. Mohammed. "Explainable AI and vision transformers for detection and classification of brain tumor: a comprehensive survey." Artificial Intelligence Review 58, no. 9 (2025): 259.

[15] Mahanty, Chandrakanta, Chin-Shiuh Shieh, Mong-Fong Horng, Anusha Nallamalla, S. Patro, Shafat Khan, and Trmesgen Engida. "Explainable vision transformer framework for multi-class classification and prognostic interpretation of oral cancer in histopathology images." Discover Oncology (2026).