QuantFormer: A Hybrid Quantum–Classical Transformer for Hyperspectral Image Classification

Jay Vinit Lunia, Saad B. Ahmed*

Department of Computer Science,
Lakehead University, Thunder Bay, ON, Canada

Abstract

Hyperspectral image (HSI) classification is challenging because each pixel has hundreds of spectral bands while only a small number of labelled samples are available. This paper presents QuantFormer, a hybrid quantum–classical transformer that embeds a small variational quantum circuit as a spectral token encoder inside a vision transformer backbone for pixel-wise land-cover mapping. A unified patch-based pipeline with band-wise normalisation, principal component analysis, and quantum token encoding is evaluated on four benchmarks: Indian Pines, Pavia University, a 7-class Houston 2013 subset, and EuroSAT_MS. With roughly 35k trainable parameters, QuantFormer attains overall accuracy above 99% on the three airborne hyperspectral datasets and about 89.8% on EuroSAT_MS, competitive with deep 3D CNNs while using substantially fewer weights. Beyond full-data experiments, we also study limited-label regimes and provide practical guidance on when quantum token encoders are a viable alternative to classical projections, without claiming quantum advantage over the strongest classical baselines.

Keywords: Hyperspectral image classification, quantum neural network, transformer, variational quantum circuit, quantum token encoder, remote sensing.

1. Introduction

Hyperspectral imaging records reflected energy in dozens to hundreds of narrow and contiguous spectral bands, forming a three dimensional cube in which each spatial pixel is associated with a high-dimensional spectrum [1]. Hyperspectral image (HSI) classification assigns a land cover label to each pixel and supports applications in agriculture, environmental monitoring, urban mapping, and mineral exploration [2]. In practice, only a small number of labeled pixels are available because field surveys and manual annotation are expensive, so models must cope with the Hughes phenomenon: overfitting caused by high spectral dimensionality and limited training data [3].

Deep spectral–spatial networks have become the dominant approach to HSI classification. Early work relied on support vector machines and hand crafted spectral features [4]. Subsequent models introduced three dimensional convolutional neural networks, spectral–spatial residual networks, and hybrid 3D–2D CNNs such as HybridSN to better capture joint spectral and spatial structure [5–7]. More recently, transformer-based architectures and state-space models have been proposed to model long-range dependencies in the spectral and spatial dimensions [8–13]. These methods achieve excellent accuracy but remain purely classical, require tens or hundreds of thousands of trainable parameters, and do not explore quantum feature maps or strict spectral information bottlenecks.

In parallel, quantum machine learning (QML) has introduced parameterised quantum circuits as trainable feature maps acting on quantum states [14–16]. Quantum classifiers and hybrid quantum classical convolutional networks have been demonstrated on small image datasets and on multispectral remote sensing benchmarks such as EuroSAT, often achieving competitive accuracy with fewer trainable parameters [17–26]. However, most

jlunia@lakeheadu.ca, *sbinahm@lakeheadu.ca

existing QML work treats quantum circuits as stand alone classifiers on low dimensional inputs, rather than as token encoders embedded in modern transformer style hyperspectral architectures.

This work examines the feasibility of replacing classical token projections with a compact quantum encoder in transformer-based HSI classifiers without sacrificing accuracy or efficiency under practical training conditions. To this end, we introduce QuantFormer, a hybrid quantum classical transformer that inserts a small variational quantum circuit as a spectral token encoder inside a vision transformer style pipeline for pixel wise HSI classification. QuantFormer is evaluated on four benchmark datasets i.e., Indian Pines, Pavia University, a 7-class Houston 2013 subset, and EuroSAT_MS under a unified training protocol, with careful comparisons between quantum token encoders and similar sized classical projections.

This work contributes to propose a transformer-based hyperspectral image (HSI) classification by introducing and rigorously evaluating a quantum token encoding mechanism based on variational quantum circuits. It demonstrates that a compact quantum encoder can be integrated as a stable tokenization module within an end-to-end trained transformer architecture on standard HSI benchmarks. The analysis quantitatively characterizes the effects of qubit count and circuit depth on classification accuracy and computational runtime, explicitly identifying saturation points beyond which additional quantum expressivity yields negligible gains. The proposed approach is further assessed under constrained learning conditions, including limited labeled data and fixed parameter budgets, establishing conditions under which quantum token encoding matches or surpasses classical projection layers of comparable complexity. Moreover, the study delineates the practical constraints associated with deploying the QuantFormer architecture on near term quantum hardware, with a detailed examination of shot noise, gradient estimation overhead, and hardware connectivity limitations that directly impact scalability and performance.

The remainder of the paper is organized into various sections. Section 2 reviews related work on classical HSI classification, transformer-based and the state space models, and quantum machine learning for imagery and remote sensing. Section 3 details the QuantFormer architecture, including preprocessing, quantum token encoding, and the transformer backbone. Section 4 summarizes the datasets, problem formulation, and evaluation metrics. Section 5 presents ablation studies. Section 6 discusses implications and hardware considerations, and Section 7 concludes the presented work with possible future direction.

2. Related Work

Early HSI methods treated each pixel independently and applied shallow classifiers such as support vector machines with spectral features or indices [4]. Deep models then introduced 3D CNNs, spectral–spatial residual networks, and hybrid 3D–2D CNNs such as HybridSN to jointly exploit spectral and spatial context and achieve near-saturated accuracy on benchmarks including Indian Pines and Pavia University [5–7]. These convolutional architectures, however, remain fully classical, require many parameters as patch size and channel dimensionality grow, and do not enforce an explicit bottleneck on spectral information.

Transformers and Sequence Models for HSI: Inspired by the success of transformers in language and RGB vision [8], recent work has adapted self-attention and state-space models to hyperspectral data. SpectralFormer and SSFTT use attention to model long-range spectral and spatial dependencies, while Mamba-style architectures treat spectral signatures as continuous-time sequences with efficient long-range interactions [9–13]. These methods provide strong classical baselines but rely on learned projections in Euclidean space and have not been combined with quantum feature maps for HSI.

Quantum Machine Learning for Imagery and Remote Sensing. QML leverages parameterised quantum circuits as expressive, potentially parameter-efficient feature maps [14–16]. Several works attach small quantum circuits to classical convolutional front-ends for digits and simple images, showing that quantum-enhanced classifiers can match CNN accuracy with fewer trainable parameters [17–20]. For remote sensing applications, hybrid quantum CNNs, quantum kernels, and quantum SVMs have been studied mainly on EuroSAT and related multispectral datasets with relatively shallow backbones [21–27]. Overall, these studies rarely consider hyperspectral cubes, patch-based pipelines, or integration with transformer-style architectures.

Hybrid Quantum Transformers. Beyond convolutional hybrids, quantum transformer architectures introduce quantum feature maps or attention mechanisms into transformer blocks. The Molecular Quantum Transformer, for example, embeds quantum circuits within attention to model electronic-structure information in molecular graphs [28]. These models target compact, structured inputs rather than large spectral cubes, and there is very limited prior work on using quantum circuits as token encoders within transformers for hyperspectral imagery.

QuantFormer lies at the intersection of these research lines. It adopts a transformer-style backbone similar in spirit to SpectralFormer and SSFTT but replaces the classical token projection with a small variational quantum circuit. Unlike many existing QML approaches that act as stand-alone classifiers on global image features, QuantFormer integrates the quantum module into a standard spectral–spatial pipeline with PCA, patch extraction, and transformer encoding, enabling a controlled comparison between quantum and classical token encoders under a unified protocol for patch-based HSI classification.

3. Methodology

This section describes the complete QuantFormer pipeline, from preprocessing and tokenisation to the quantum encoder, transformer backbone, and training procedure.

Preprocessing and Tokenization: Let, $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote an HSI cube with height H, width W, and C spectral bands. The cube is first converted to floating-point values and normalised on a per-band basis using min–max scaling. Let $\mathbf{X}_{\min}, \mathbf{X}_{\max} \in \mathbb{R}^{1\times 1\times C}$ denote the per-band minima and maxima computed over the training portion of the cube. The normalised data are

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X} - \mathbf{X}_{\min}}{\mathbf{X}_{\max} - \mathbf{X}_{\min} + \varepsilon}, \tag{3.1}$$

with $\varepsilon = 10^{-6}$ to avoid division by zero. This mapping keeps each band in $[0, 1]$ and prevents very large or small values from generating unstable quantum rotation angles.

To reduce spectral dimensionality and mitigate the Hughes phenomenon, the pipeline applies PCA along the spectral dimension [3]. Let $\mathbf{\Sigma} \in \mathbb{R}^{C\times C}$ be the covariance matrix of the normalised spectra, and let $\{\lambda_i, \mathbf{u}_i\}_{i=1}^{C}$ be its eigenvalue–eigenvector pairs ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_C \geq 0$. The number of retained components k is chosen to satisfy

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{C} \lambda_i} \geq 0.99, \tag{3.2}$$

so that at least 99% of the spectral variance is kept. The projection matrix is

$$\mathbf{W}_k = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k] \in \mathbb{R}^{C\times k}, \tag{3.3}$$

and the PCA-compressed cube is

$$\mathbf{X}_{\text{PCA}} = \mathbf{X}_{\text{norm}} \mathbf{W}_k. \tag{3.4}$$

For the datasets considered, this reduces the number of channels to roughly $k \in [15, 30]$ while preserving most discriminative information.

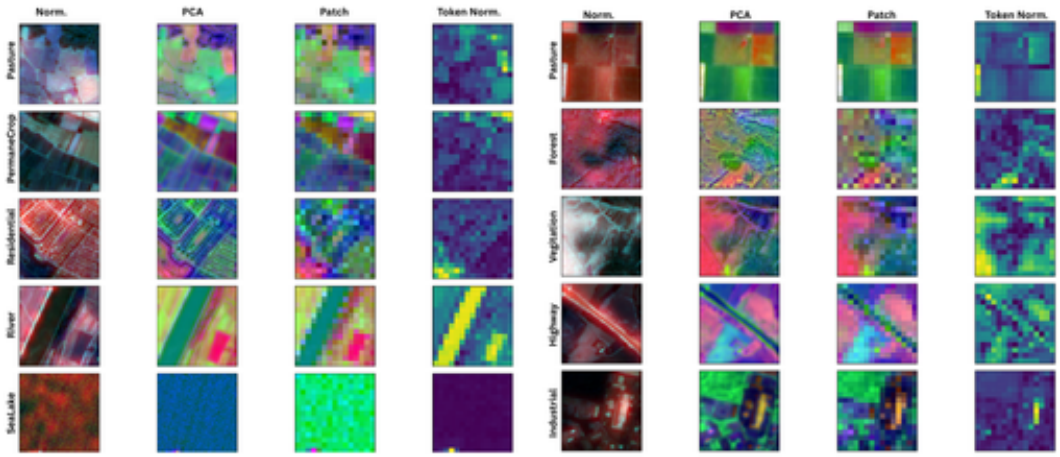

Figure 1. EuroSAT_MS preprocessing pipeline. For each class, one example chip is shown at four stages: per-band min–max normalisation, PCA-compressed cube, 15×15 centre patch, and token-wise ℓ_2-norm heatmap over the patch.

For each labelled pixel at spatial coordinates (i, j), a square patch of size $p \times p$ is extracted from $\mathbf{X}_{\text{PCA}}$:

$$\mathbf{P}_{i,j} \in \mathbb{R}^{p \times p \times k}. \tag{3.5}$$

Zero-padding is applied at the borders so that all labelled pixels have a valid patch. If the label map uses a background index 0, only pixels with $y_{i,j} \neq 0$ are considered.

Each patch is then reshaped into a sequence of spectral tokens. Let p be the patch size and $N = p^2$ the number of spatial positions in a patch. A patch $\mathbf{P}_{i,j}$ is permuted to shape (k, p, p) and flattened across spatial dimensions:

$$\mathbf{T}_{i,j} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N]^\top \in \mathbb{R}^{N \times k}, \tag{3.6}$$

where each token $\mathbf{t}_n \in \mathbb{R}^k$ is the PCA-compressed spectrum at one location in the patch. In practice, $p = 15$, so each patch yields $N = 225$ tokens.

Figure 1 illustrates the preprocessing stages on EuroSAT_MS, from per-band normalisation and PCA compression to patch extraction and token-wise ℓ_2-norm maps.

Quantum Token Encoder: The quantum token encoder maps each PCA-compressed token $\mathbf{t}_n \in \mathbb{R}^k$ into a quantum-enhanced feature vector using a parameterised quantum circuit. The encoder has three steps: computing rotation angles, applying angle embedding and entangling layers, and measuring Pauli-Z expectations. Figure 2 illustrates this encoder for EuroSAT_MS.

First, a classical linear layer maps each token to n_q rotation angles:

$$\boldsymbol{\theta}_n = \pi \cdot \tanh(\mathbf{W}_{\text{angle}} \mathbf{t}_n + \mathbf{b}_{\text{angle}}), \tag{3.7}$$

where $\mathbf{W}_{\text{angle}} \in \mathbb{R}^{n_q \times k}$ and $\mathbf{b}_{\text{angle}} \in \mathbb{R}^{n_q}$ are trainable parameters. The tanh function keeps angles in $[-\pi, \pi]$.

The quantum circuit operates on n_q qubits initialised in the ground state $|0\rangle^{\otimes n_q}$. Angle embedding applies single-qubit rotations around the Y-axis:

$$R_y(\theta) = \begin{bmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{bmatrix}, \tag{3.8}$$

so that after embedding

$$|\psi_{\text{emb}}(\boldsymbol{\theta}_n)\rangle = \bigotimes_{q=1}^{n_q} R_y(\theta_{n,q})|0\rangle. \tag{3.9}$$

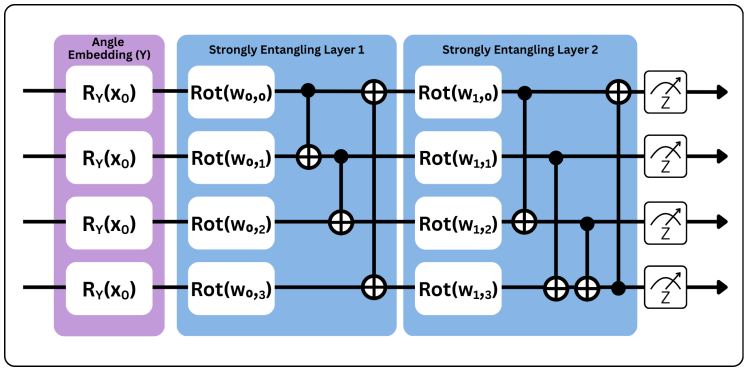

Figure 2. Illustration of the quantum token encoder on EuroSAT_MS. PCA-compressed spectral tokens are processed by the variational circuit, and the resulting embeddings are projected back to a spatial map to visualise class structure.

Several strongly entangling layers follow, composed of parameterised single-qubit rotations and controlled-NOT gates in the pattern used by the `StronglyEntanglingLayers` template in PennyLane [29]. If $\boldsymbol{\phi}$ denotes all circuit parameters, the final state is

$$|\psi_{\text{enc}}(\boldsymbol{\theta}_n, \boldsymbol{\phi})\rangle = U_{\text{ent}}(\boldsymbol{\phi})|\psi_{\text{emb}}(\boldsymbol{\theta}_n)\rangle. \tag{3.10}$$

The circuit output is obtained by measuring the expectation value of the Pauli-Z operator on each qubit:

$$z_{n,q} = \langle\psi_{\text{enc}}(\boldsymbol{\theta}_n, \boldsymbol{\phi})|Z_q|\psi_{\text{enc}}(\boldsymbol{\theta}_n, \boldsymbol{\phi})\rangle, \tag{3.11}$$

where Z_q acts as Pauli-Z on qubit q. The resulting vector of quantum features is

$$\mathbf{z}_n = \left[z_{n,1}, z_{n,2}, \ldots, z_{n,n_q}\right]^\top \in \mathbb{R}^{n_q}. \tag{3.12}$$

Athough mapping k spectral components to $n_q = 4$ qubits compresses the input, the trainable parameters $\boldsymbol{\phi}$ and the entanglement pattern in U_{ent} allow the circuit to implement a non-linear kernel that emphasises class-discriminative spectral manifolds before measurement. A final classical linear layer maps $\mathbf{z}_n$ to the transformer model dimension d_{model}:

$$\mathbf{h}_n = \mathbf{W}_{\text{proj}}\mathbf{z}_n + \mathbf{b}_{\text{proj}}, \tag{3.13}$$

with $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_{\text{model}} \times n_q}$ and $\mathbf{b}_{\text{proj}} \in \mathbb{R}^{d_{\text{model}}}$. In this work, $n_q = 4$ and $d_{\text{model}} = 64$.

This design can be interpreted as implementing a trainable quantum feature map followed by a classical linear projection [15]. The ablation studies in Section 5 show that reducing the encoder to a single qubit significantly harms accuracy, while a four-qubit encoder remains competitive with purely classical projections of similar size.

Transformer Backbone and Classifier: The sequence of quantum-enhanced token embeddings is processed by a shallow transformer encoder. Let $\mathbf{H}_0 \in \mathbb{R}^{N \times d_{\text{model}}}$ denote the matrix whose rows are the token embeddings $\mathbf{h}_n$. Fixed sinusoidal positional encodings are added to inject spatial information [8]. For position index $\text{pos} \in \{0, \ldots, N-1\}$ and dimension index i,

$$\text{PE}_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right), \tag{3.14}$$

$$\text{PE}_{(\text{pos},2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right), \tag{3.15}$$

and the final input to the transformer encoder is

$$\mathbf{H}_0^{\text{PE}} = \mathbf{H}_0 + \mathbf{PE}, \tag{3.16}$$

where $\mathbf{PE} \in \mathbb{R}^{N \times d_{\text{model}}}$ stacks all positional vectors.

QuantFormer uses a single encoder layer with model dimension $d_{\text{model}} = 64$, 4 attention heads, and feed-forward dimension $d_{\text{ff}} = 128$ [8]. Let $\mathbf{H}^{(0)} = \mathbf{H}_0^{\text{PE}}$. For layer index ℓ,

$$\tilde{\mathbf{H}}^{(\ell)} = \mathbf{H}^{(\ell-1)} + \text{MHSA}\Big(\text{LN}\Big(\mathbf{H}^{(\ell-1)}\Big)\Big), \tag{3.17}$$

$$\mathbf{H}^{(\ell)} = \tilde{\mathbf{H}}^{(\ell)} + \text{FFN}\Big(\text{LN}\Big(\tilde{\mathbf{H}}^{(\ell)}\Big)\Big), \tag{3.18}$$

where LN denotes layer normalisation, MHSA is multi-head self-attention, and FFN is a position-wise feed-forward network. The attention mechanism is

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3.19}$$

with d_k the head dimension. The feed-forward network has the form

$$\text{FFN}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\,\mathbf{W}_2 + \mathbf{b}_2. \tag{3.20}$$

Empirically, a single layer is sufficient to reach near-saturated accuracy on the considered datasets once the quantum encoder has enriched the token representations.

After the transformer, token features are aggregated via global average pooling. Let $\mathbf{H}^{(L)} \in \mathbb{R}^{N \times d_{\text{model}}}$ be the output of the final layer. The pooled feature is

$$\mathbf{h}_{\text{pool}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}^{(L)}_{n,:}, \tag{3.21}$$

and the classifier maps $\mathbf{h}_{\text{pool}}$ to logits

$$\mathbf{o} = \mathbf{W}_{\text{cls}}\mathbf{h}_{\text{pool}} + \mathbf{b}_{\text{cls}}, \tag{3.22}$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{C_{\text{cls}} \times d_{\text{model}}}$ and $\mathbf{b}_{\text{cls}} \in \mathbb{R}^{C_{\text{cls}}}$. The predicted probability for class c is

$$\hat{y}_c = \frac{\exp(o_c)}{\sum_{j=1}^{C_{\text{cls}}} \exp(o_j)}. \tag{3.23}$$

Loss Function and Training: Let, $\mathbf{P}_n$ be a patch and y_n its ground-truth class. The model predicts $\hat{\mathbf{y}}_n = f_{\boldsymbol{\theta}}(\mathbf{P}_n)$, where $\boldsymbol{\theta}$ collects all classical and quantum parameters. The main loss term is cross-entropy between $\hat{\mathbf{y}}_n$ and a one-hot encoding of y_n, summed over all samples. Cross-entropy is standard in HSI classification and produces smooth gradients and calibrated probabilities [4–7]. An ℓ_2 regularisation term (weight decay) is applied to the classical weights in the angle-projection, transformer, and classifier layers [2]. The quantum circuit is regularised implicitly by its shallow depth and small qubit count.

The classical parts of the network are implemented in PyTorch, and the quantum circuit is implemented with PennyLane as a differentiable state-vector simulator [29]. Training uses the Adam optimiser with learning rate 2×10^{-3} and weight decay 10^{-4}, together with global gradient clipping to prevent gradient explosions. Mini-batches of patches and labels are sampled, passed through the entire pipeline, and the loss is minimised via backpropagation through both classical and quantum components using analytic gradients [30]. Batch sizes of 32 for training and 64 for evaluation and 50 training epochs per dataset give a good trade-off between stability and runtime on CPU-based simulators. All experiments were executed on an Apple MacBook Pro with an Apple Silicon M4 chip (CPU-only simulation; no external GPU).

Table 1 summarises the main hyperparameters used for all experiments.

Figure 3 summarizes the overall data flow, from the original HSI cube to the final class probabilities.

Table 1. Hyperparameter settings for QuantFormer.

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	2×10^{-3}
Weight decay	10^{-4}
Batch size (train / eval)	32/64
Epochs per dataset	50
Patch size p	15
PCA variance threshold	99%
Qubits n_q	4
Model dimension d_{model}	64
Feed-forward dimension d_{ff}	128
Attention heads	4

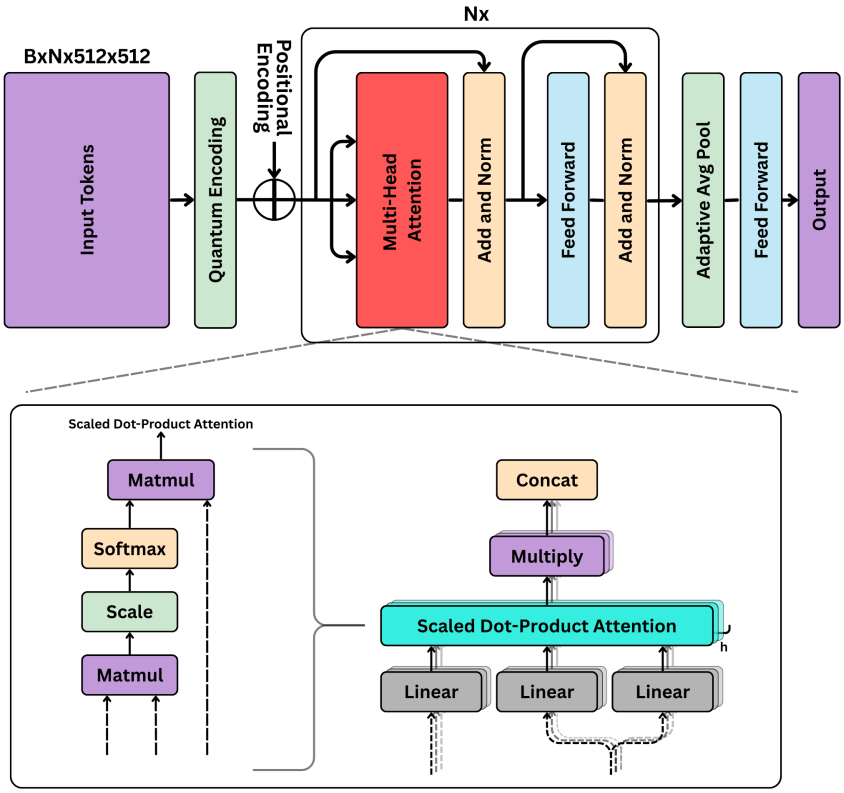

Figure 3. QuantFormer Architecture: PCA compresses the spectral dimension of the HSI cube, local patches are tokenised, tokens are mapped into a quantum Hilbert space by a variational quantum circuit (VQC), and the resulting quantum-enhanced embeddings are processed by a transformer encoder before global pooling and classification.

4. Dataset Description

This section summarizes the details of the datasets. Benchmark datasets such as Indian Pines, Pavia University, and Houston 2013 are airborne hyperspectral datasets with hundreds or tens of spectral bands, while EuroSAT_MS is a multispectral satellite dataset derived from Sentinel-2 imagery and adapted here to match the patch based HSI protocol.

Indian Pines is an agricultural scene captured by the AVIRIS sensor over north western Indiana [31]. After removing noisy and water absorption bands, the cube has size $145 \times 145 \times 200$ with 16 land cover classes such as corn, soybeans, and woodland.

Pavia University is an urban scene acquired by the ROSIS sensor over the University of Pavia, Italy [32]. The cube has spatial dimensions 610×340 with $C = 103$ spectral bands and 9 land cover classes including asphalt, meadows, trees, and several man-made structures.

Houston 2013 is an airborne hyperspectral dataset from the 2013 IEEE GRSS data fusion contest [33]. The subset used here has size 954×210 with $C = 48$ spectral channels and 7 aggregated urban and vegetation classes.

EuroSAT_MS is a satellite dataset derived from ESA Sentinel-2 multispectral imagery for land-use and land-cover classification [34]. Each sample is a 64×64 chip with $C = 13$ spectral bands and 10 classes. A 15×15 centre patch is used as the model input for each chip to align the input geometry with the HSI patch protocol; the reported EuroSAT_MS results are therefore not directly comparable to full-chip EuroSAT benchmarks.

Let, $\mathbf{Y} \in \{0, 1, \ldots, C_{\text{cls}} - 1\}^{H \times W}$ be the label map for each cube, where C_{cls} is the number of land cover classes and label 0 denotes background when present. The model learns a function that maps a patch centered at a labeled pixel to a distribution over classes.

$$f_{\boldsymbol{\theta}} : \mathbb{R}^{p \times p \times C} \to \Delta^{C_{\text{cls}}-1}, \tag{4.1}$$

Where, $\Delta^{C_{\text{cls}}-1}$ is the probability simplex. Given a dataset,

$$\mathcal{D} = \{(\mathbf{P}_n, y_n)\}_{n=1}^{N}, \tag{4.2}$$

with patches $\mathbf{P}_n$ and labels $y_n \in \{0, \ldots, C_{\text{cls}} - 1\}$, the empirical loss

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{P}_n), y_n) \tag{4.3}$$

is minimized, where $\mathcal{L}$ is the total loss defined in Section 3.

Performance is reported using overall accuracy (OA), average accuracy (AA), and Cohen's κ coefficient. OA measures the proportion of correctly classified pixels, AA averages per-class accuracies to reduce the effect of class imbalance, and κ measures agreement beyond chance based on the confusion matrix. Training and inference times, together with the number of trainable parameters, are also reported to characterise model efficiency.

5. **Ablation Studies**

To better understand the contribution of each component as specified in Section 3, several ablation variants are evaluated: (i) removal of the quantum encoder in favor of a classical linear projection (w/o QNN), (ii) removal of positional encodings (w/o PosEnc), and (iii) reduction of the quantum encoder to a single qubit (1 Qubit). All variants share the same preprocessing, transformer configuration, optimizer, and dataset protocol.

All experiments use PyTorch for the classical components and PennyLane for the quantum simulator [29]. The quantum device is a noiseless state-vector simulator with $n_q = 4$ wires and analytic gradients. All runtimes are measured on the same Apple MacBook Pro with Apple Silicon M4 (CPU-only execution; no external GPU); absolute values depend on hardware, but relative trends across configurations are informative.

Table 2 reports the ablation results on Indian Pines. Removing the quantum encoder slightly improves OA in this near-saturated setting, while dropping positional encodings or using a single qubit clearly hurts accuracy, showing that spatial information and several qubits are still important.

Table 2. Ablation study on Indian Pines.

Config.	**OA(%)**	**AA(%)**	κ	**Params(k)**	**Train(s)**	**Infer(ms)**
Full	99.08	98.86	0.990	35.2	826	0.43
w/o QNN	99.60	98.12	0.995	36.4	377	0.13
w/o PosEnc	98.31	96.50	0.981	35.2	837	0.43
1 Qubit	97.28	94.58	0.969	34.9	410	0.16

Table 3. Ablation study on Pavia University.

Config.	**OA(%)**	**AA(%)**	κ	**Params(k)**	**Train(s)**	**Infer(ms)**
Full	99.94	99.92	0.999	34.7	7278	0.90
w/o QNN	99.94	99.94	0.999	34.6	2330	0.25
w/o PosEnc	99.24	98.68	0.990	34.7	7238	0.78
1 Qubit	97.84	96.37	0.971	34.5	2913	0.34

Table 4. Ablation study on Houston 2013.

Config.	**OA(%)**	**AA(%)**	κ	**Params(k)**	**Train(s)**	**Infer(ms)**
Full	99.92	99.93	0.999	34.5	202	0.43
w/o QNN	99.92	99.93	0.999	34.4	91	0.12
w/o PosEnc	99.88	99.87	0.999	34.5	203	0.42
1 Qubit	95.38	95.48	0.946	34.3	100	0.16

Table 5. Overall performance of QuantFormer (single seed per dataset).

Dataset	**OA(%)**	**AA(%)**	κ	**Params(k)**	**Train(s)**	**Infer(ms)**
Indian Pines	99.08	98.86	0.990	35.2	826	0.43
Pavia Univ.	99.94	99.92	0.999	34.7	7278	0.90
Houston 2013	99.92	99.93	0.999	34.5	202	0.43
EuroSAT_MS	89.83	89.62	0.887	34.8	2233	0.98

Table 3 shows that, on Pavia University, all high-capacity variants reach very high accuracy. The quantum encoder matches the classical projection with a slightly smaller parameter count, whereas the single-qubit encoder degrades performance.

z

Table 4 shows that Houston 2013 is almost linearly separable after PCA, with all high-capacity variants saturating near-perfect accuracy. In this regime the quantum encoder mainly affects runtime, while the single-qubit variant still reduces accuracy; the confusion matrix for the full model, corresponding to the "Full" row, is shown in the third panel of Figure 4.

Table 5 summarizes overall performance and efficiency. QuantFormer achieves OA above 99% on the three airborne hyperspectral datasets with about 35k trainable parameters and sub-millisecond per-sample inference on the simulator. On EuroSAT_MS, the model reaches OA of about 89.8%, which is lower than the airborne benchmarks because only a 15×15 centre patch from each 64×64 chip is used and the dataset provides coarser spectral resolution with 13 bands, reducing available spatial and spectral context. Overall, these results indicate that the combination of PCA, patch-based tokenization, quantum token encoding, and a shallow transformer encoder can match the accuracy of strong classical models at a modest parameter cost. Figure 4 shows the confusion matrices for the full

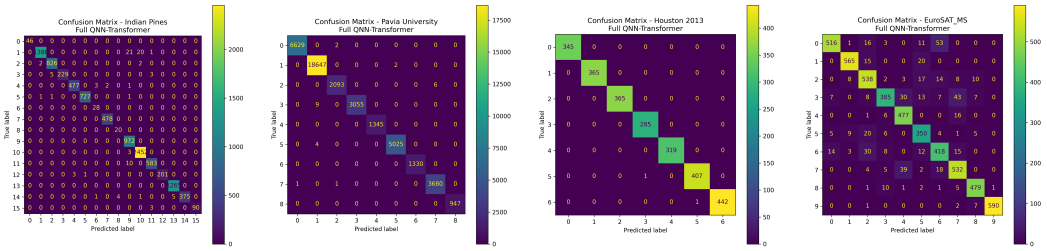

Figure 4. Confusion matrices for the full QuantFormer configuration on (from left to right) Indian Pines, Pavia University, Houston 2013, and EuroSAT_MS.

Table 6. Classification results on Indian Pines (IP) and Pavia University (PU). Baseline values are taken from the original papers. A dash indicates that a result was not reported for that dataset.

	IP			PU		
Method	**OA(%)**	**AA(%)**	κ	**OA(%)**	**AA(%)**	κ
SVM [4]	85.30	79.03	0.831	94.34	92.98	0.925
3D-CNN [5]	91.10	91.58	0.900	—	—	—
SSRN [6]	99.19	98.93	0.991	99.90	99.91	0.999
HybridSN [7]	99.75	99.63	0.997	99.98	99.97	0.999
SpectralFormer [9]	93.45	91.28	0.926	96.46	95.77	0.953
SSFTT [10]	98.72	97.86	0.986	99.52	99.23	0.994
MorphMamba [11]	99.48	98.76	0.994	—	—	—
SS-Mamba [13]	—	—	—	99.94	99.89	0.999
QuantFormer	**99.08**	**98.86**	**0.990**	**99.94**	**99.92**	**0.999**

QuantFormer configuration on all four datasets, with Houston 2013 corresponding to the third panel.

6. Discussion

The experimental results show that QuantFormer is capable of reaching very high accuracy on three airborne hyperspectral benchmarks and competitive accuracy on EuroSAT_MS while using 35k trainable parameters. A unified preprocessing pipeline combined with patch-based tokenization allows the method to generalize across different HSI cubes without requiring substantial dataset-specific customization. Table 6 places these results in the context of existing methods on Indian Pines and Pavia University. Baselines include SVMs, 3D CNNs, spectral-spatial residual networks, HybridSN, transformer-based models, and recent state-space architectures [4–7, 9–13]. Unless stated otherwise, the baseline numbers are taken from the original publications and may use slightly different training protocols.

On Indian Pines and Pavia University, QuantFormer reaches the same accuracy regime as the strongest convolutional and transformer-based baselines while using a much smaller transformer backbone. It clearly improves over shallow methods such as SVMs and early 3D CNNs, indicating that replacing the classical token projection with a compact quantum encoder preserves competitive accuracy at modest parameter cost.

All quantum circuits are run on a noiseless state-vector simulator with analytic gradients, so the reported numbers do not capture noise, decoherence, or limited qubit connectivity on current hardware. Several baselines in Table 6 are taken from the literature review and may use slightly different training protocols or data augmentations, so QuantFormer should be viewed as a compact and competitive hybrid design rather than a claimed state-of-the-art

model. The four-qubit circuit fits within many near-term devices, but deploying QuantFormer on real hardware will require estimating gradients with finite shots and compiling the entangling pattern to hardware-native two-qubit gates, which may increase depth and training cost [30].

Overall, the results indicate that hybrid quantum classical transformers with compact token encoders are a promising direction for hyperspectral image classification when model size and label budgets are constrained, and they provide a foundation for more hardware focused and semi-supervised work in future studies.

7. Conclusion

This paper has presented QuantFormer, a hybrid quantum classical transformer that combines simple preprocessing, PCA-based spectral compression, patch-based tokenization, a compact variational quantum circuit for token encoding, and a shallow transformer encoder for hyperspectral image classification. Experiments on Indian Pines, Pavia University, Houston 2013, and EuroSAT_MS show that QuantFormer attains OA above 99% on three airborne hyperspectral datasets and around 89.8% on EuroSAT_MS with about 35k parameters, yielding a compact spectral representation that matches strong classical baselines while enforcing a strict information bottleneck on the spectral features.

Future work includes executing QuantFormer on cloud-based quantum hardware to compare simulator and hardware performance, exploring deeper or dataset-adaptive quantum encoders and quantum attention mechanisms, and investigating semi-supervised training on large unlabeled HSI archives and higher-resolution UAV and satellite datasets.

References

[1] C. Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Springer, 2003.

[2] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson. "Advances in Hyperspectral Image Classification: Earth Monitoring with Statistical Learning Methods". In: *IEEE Signal Processing Magazine* 31.1 (2014), pp. 45–54.

[3] C. Rodarmel and J. Shan. "Principal component analysis for hyperspectral image classification". In: *Surveying and Land Information Science* (2002).

[4] F. Melgani and L. Bruzzone. "Classification of hyperspectral remote sensing images with support vector machines". In: *IEEE Transactions on Geoscience and Remote Sensing* 42.8 (2004), pp. 1778–1790.

[5] A. Ben Hamida, A. Benoit, P. Lambert, and C. Ben Amar. "3-D Deep Learning Approach for Remote Sensing Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 56.8 (2018), pp. 4420–4434.

[6] Z. Zhong, J. Li, J. Luo, and M. Chapman. "Spectral–Spatial Residual Network for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 56.2 (2018), pp. 897–909.

[7] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri. "HybridSN: Exploring 3D–2D CNN Feature Hierarchy for Hyperspectral Image Classification". In: *IEEE Geoscience and Remote Sensing Letters* 17.2 (2020), pp. 277–281.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. 2017.

[9] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot. "SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022).

[10] L. Sun, G. Zhao, Y. Zheng, and Z. Wu. "Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022).

[11] M. H. Butt, S. Li, and A. Plaza. "Spatial–Spectral Morphological Mamba for Hyperspectral Image Classification". In: *Neurocomputing* (2024).

[12] R. Khan, T. Arshad, X. Ma, H. Zhu, C. Wang, J. Khan, Z. U. Khan, and S. U. Khan. "GroupFormer for Hyperspectral Image Classification Through Group Attention". In: *Scientific Reports* 14 (2024).

[13] H. Zhang, X. Xu, S. Li, and A. Plaza. "Wavelet Decomposition-Based Spectral–Spatial Mamba Network for Hyperspectral Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[14] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe. "Circuit-Centric Quantum Classifiers". In: *Physical Review A* 101.3 (2020).

[15] M. Schuld. "Supervised Quantum Machine Learning Models Are Kernel Methods". In: *arXiv preprint arXiv:2101.11020* (2021).

[16] A. Perez-Salinas. "Data re-uploading for a universal quantum classifier". In: *Quantum* 4 (2020), p. 226.

[17] A. Senokosov. "Quantum machine learning for image classification". In: *Machine Learning: Science and Technology* (2023).

[18] F. Riaz, S. Abdulla, H. Suzuki, S. Ganguly, R. C. Deo, and S. Hopkins. "Accurate image multi-class classification neural network model with quantum computing". In: *Sensors* 23.5 (2023).

[19] M. Henderson, S. Shakya, S. Pradhan, and T. Cook. "Quanvolutional neural networks: Powering image recognition with quantum circuits". In: *Quantum Machine Intelligence* 2 (2020).

[20] J. Liu. "Hybrid quantum-classical convolutional neural networks". In: *Science China Physics, Mechanics and Astronomy* 64 (2021).

[21] A. Sebastianelli, D. A. Zaidenberg, D. Spiller, B. Le Saux, and S. Ullo. "On Circuit-Based Hybrid Quantum Neural Networks for Remote Sensing Imagery Classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022).

[22] S. Otgonbaatar and M. Datcu. "Classification of remote sensing images with parameterized quantum circuits". In: *IEEE Geoscience and Remote Sensing Letters* 19 (2022).

[23] Y. Liu. "Quantum machine learning on remote sensing data classification". In: *Journal of Engineering Research Studies* 2.12 (2022).

[24] ESA-PhiLab. *QNN4EO: Quantum convolutional neural network for satellite data classification*. GitHub repository. 2021.

[25] M. Zaman, T. Kehkashan, A. Akhunzada, H. Alaidaros, M. Uddin, and M. Azeem. "EQCNN: Enhanced Remote Sensing Imagery Classification with Circuit-Based Error-Corrected Quantum Convolutional Neural Networks". In: *Proc. DICTA*. 2024.

[26] H. Kumar. "Remote sensing classification using quantum image processing". In: *Proceedings of SPIE*. Vol. 13196. 2024.

[27] D. Mazur, T. Rybotycki, and P. Gawron. *Hyperspectral Image Segmentation with a Machine Learning Model Trained Using Quantum Annealer*. arXiv:2503.01400. 2025.

[28] Y. Kamata, Q. H. Tran, Y. Endo, and H. Oshima. "Molecular Quantum Transformer". In: *arXiv preprint arXiv:2503.21686* (2025).

[29] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, and N. Killoran. "PennyLane: Automatic Differentiation of Hybrid Quantum-Classical Computations". In: *arXiv preprint* (2018). arXiv: 1811.04968 [quant-ph].

[30] M. Schuld, V. Bergholm, C. Gogolin, J. Izaac, and N. Killoran. "Evaluating Analytic Gradients on Quantum Hardware". In: *Physical Review A* 99.3 (2019), p. 032331.

[31] M. F. Baumgardner. *220 Band AVIRIS Hyperspectral Image Data Set*. Purdue University Research Repository. 2015.

[32] Pavia University Dataset. *Hyperspectral Image Data*. http://www.ehu.eus/ccwintco/. 2003.

[33] C. Debes. "Hyperspectral and LiDAR data fusion". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7.6 (2014).

[34] P. Helber, B. Bischke, A. Dengel, and D. Borth. "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.7 (2019), pp. 2217–2226.