RESEARCH PAPER

Quantum-enhanced convolutional networks for animal image classification in environmental monitoring: a comparison with Vision Transformer and VGG16

Banavathu Prasanth Kumar[a], Vithya Ganesan[a,*]

[a] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India*

* Corresponding author: prasanth050802@gmail.com Co-author e-mail: vithyamtech@gmail.com.

Abstract

Automated recognition of animals in images supports wildlife surveys, livestock monitoring and other environmental applications, and it needs classifiers that stay accurate on varied images. This study compares three deep learning models for image classification on the CIFAR-10 benchmark: a quantum-enhanced convolutional neural network (EQ-CNN), a Vision Transformer (ViT) and VGG16. CIFAR-10 contains six animal classes (bird, cat, deer, dog, frog, horse) and four vehicle classes, so it allows animal and non-animal classes to be compared under identical conditions. All images were resized to 224 × 224 pixels and normalised, and each model was trained for five epochs. The EQ-CNN combines convolutional layers with quantum-inspired non-linear activations. ViT encodes images with self-attention, and VGG16 is the classical convolutional baseline. Performance was assessed with accuracy, precision, recall, F1-score, confusion matrices, ROC curves, RMSE and R^2. K-means clustering and class similarity graphs were used to examine feature separation. The EQ-CNN scored highest in every class; its mean per-class score over the six animal classes was 0.945, against 0.913 for ViT and 0.877 for VGG16. The results suggest that quantum-inspired hybrid designs are worth testing on real animal-monitoring data such as camera-trap images. Performance was evaluated in terms of accuracy, precision, recall, F1-score, confusion matrices, ROC curves, RMSE and R^2. The K-means clustering and class similarity graphs were used to analyze feature separation. The EQ-CNN achieved the highest mean per-class score (0.945) across all the animal classes, compared to ViT (0.913) and VGG16 (0.877). The results suggest that hybrid designs inspired by quantum should be tested on real animal monitoring data, for example on camera-trap images.

Keywords: Animal image classification; Environmental monitoring; Quantum convolutional neural network; Vision Transformer; VGG16; CIFAR-10; Deep learning

1. Introduction

Performance was evaluated in terms of accuracy, precision, recall, F1-score, confusion matrices, ROC curves, RMSE and R^2. The K-means clustering and class similarity graphs were used to analyze feature separation. The EQ-CNN achieved the highest mean per-class score (0.945) across all the animal classes, compared to ViT (0.913) and VGG16 (0.877). The results suggest that hybrid designs inspired by quantum should be tested on real animal monitoring data, for example on camera-trap images.

Traditional convolutional neural networks like VGG16 [18] are still strong feature extractors. Vision Transformers (ViT) [17] replace local convolution with attention which enables the model to relate distant parts of an image. Quantum inspired networks draw inspiration from quantum physics such as superposition, entanglement and quantum kernels to provide the network with richer non-linear mappings and better generalization [16,19,22].

We compare a CIFAR-10 [24] based VGG16 CNN, an improved quantum CNN (EQ-CNN) and a Vision Transformer. Six of its ten classes are animals, so the benchmark provides a controlled initial test of how the models treat animal images. We report accuracy, precision, recall, F1-score, confusion matrices and ROC-AUC . We analyze the class organization of each model in feature space through K-means clustering and similarity graphs.The main contributions are:

- A like-for-like comparison of a quantum-enhanced CNN, ViT and VGG16 under one preprocessing and training setup.
- A per-class analysis that reports the six animal classes separately from the four vehicle classes.
- Embedding-level analysis with K-means clustering and class similarity networks to show where each model confuses similar classes.

2. Related work

We compare a CIFAR-10 [24] based VGG16 CNN, an improved quantum CNN (EQ-CNN) and a Vision Transformer. Six of its ten classes are animals, so the benchmark provides a controlled initial test of how the models treat animal images. We report accuracy, precision, recall, F1-score, confusion matrices and ROC-AUC . We analyze the class organization of each model in feature space through K-means clustering and similarity graphs.Farhi and Neven [7] applied quantum neural networks to classical image data, and Sen and Banerjee [8] used quantum entanglement to improve transformer attention.

There are a number of studies on practical limits. Lightweight hybrid models to reduce the computing requirements were proposed by Liu et al. [9] and Mehta and Jain [14], and the impact of preprocessing on the hybrid models was studied by Dutta and Pradhan [10]. The mixing of quantum and classical methods for visual recognition was reviewed by Li and Cheng [11], Batra and Raj [12] and Singh and Ahmed [13] and Patel and Banerjee [15] combined quantum methods with vision transformers for multimodal learning.

***Corresponding Author:**

Received: 25th May, 2026; Revised: 6th July, 2026; Accepted: 28th July, 2026; Available Online: 29th August, 2026

(DOI): 10.70102/AEJ.2026.18.8s.21

The comparison builds on prior work. Cong et al. [16] proposed quantum convolutional neural networks, Dosovitskiy et al. [17] proposed the Vision Transformer and Simonyan and Zisserman [18] demonstrated the effectiveness of very deep CNNs. Biamonte et al. [19] and Abbas et al. [21] provided reviews on the potential of quantum machine learning and quantum neural networks, Han et al. [20] surveyed Vision Transformers, and Schuld and Killoran [22] described feature Hilbert spaces. The representational ability of graph neural networks was examined by Xu et al. [23], which motivated us to use similarity graphs.

3. Materials and methods

3.1. Dataset

CIFAR-10 [24] comprises ten classes of 60,000 color 32 × 32 pixel pictures. There are four classes of vehicles (plane, automobile, ship, truck) and six classes of animals (bird, cat, deer, dog, frog, horse). With an equal distribution of classes, the data was divided into 50,000 training photos and 10,000 test images. No live animals were used in this investigation, and the dataset is available to the public.

3.2. Preprocessing

Every image was transformed to tensors, normalized to a standard range, and shrunk to 224 × 224 pixels, the input size needed by VGG16 and the Vision Transformer. During training, random horizontal flips were used to enhance generalization. The labels remained unaltered. As a result, each model was given the same kind of input, maintaining the comparison's fairness.

3.3. Model architectures

The overall workflow is shown in Fig. 1. The three models are described below.

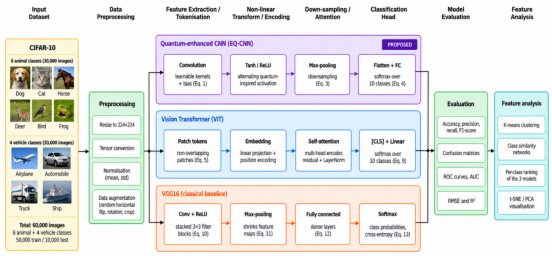


Fig. 1. Workflow of the study: input dataset, preprocessing, the three model branches (quantum-enhanced CNN, Vision Transformer, VGG16), evaluation metrics, and visualisation and comparison.

3.3.1. Quantum-enhanced CNN (EQ-CNN)

The EQ-CNN employs quantum-inspired activation patterns while maintaining a conventional convolutional structure. Wave-like, non-classical feature mappings are imitated by alternating Tanh and ReLU activations [4,16]. There was no usage of quantum hardware; these are classical processes. A bias term modifies the reaction to various data, and learnable kernels search the feature maps for local patterns like edges and textures (Eq. 1).

$$Z^{(l)} = W^{(l)} * X^{(l-1)} + b^{(l)} \tag{1}$$

Tanh maps values to the range −1 to 1 (Eq. 2). This adds non-linearity and keeps both positive and negative signals, which supports richer feature learning.

$$A^{(l)} = \tanh\left(Z^{(l)}\right) = \frac{e^{Z^{(l)}} - e^{-Z^{(l)}}}{e^{Z^{(l)}} + e^{-Z^{(l)}}} \tag{2}$$

Max-pooling (Eq. 3) reduces data size, keeps the important features and makes the network less sensitive to small shifts.

$$P^{(l)} = \text{MaxPool}\left(A^{(l)}\right) \tag{3}$$

The pooled maps are flattened and passed to a fully connected layer (Eq. 4). A softmax function then gives class probabilities for the ten CIFAR-10 classes.

$$F = \text{Flatten}\left(P^{(n)}\right), \quad Y = \text{Softmax}\left(W_{fc} \cdot F + b_{fc}\right) \tag{4}$$

3.3.2. Vision Transformer (ViT)

ViT uses a Transformer encoder to analyze the image after dividing it into small patches. Self-attention enables the model to take into account the entire scene at once, capturing the relationships between far-off areas.

Patching: each non-overlapping patch in the image is flattened into a token (Eq. 5). This allows image data to be handled as a series and may be accomplished with PyTorch unfold.

$$X = \{x_{p_1}, x_{p_2}, \ldots, x_{p_N}\}, \quad x_{p_i} \in R^{P \times P \times C} \tag{5}$$

Linear projection: each patch token is mapped to a higher-dimensional space by a learnable linear layer, and positional encodings are added to preserve spatial order (Eq. 6).

$$Z^0 = \left[x_{p_1}E; x_{p_2}E; \ldots; x_{p_N}E\right] + E_{pos} \tag{6}$$

Multi-head self-attention: each token uses scaled dot-product attention to pay attention to every other token. Linear projections are the source of queries, keys, and values, which are processed in parallel across several heads (Eq. 7).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

Transformer layer output: layer normalization stabilizes the gradients, and residual connections integrate the attention output with its input (Eq. 8).

$$Z^{(l+1)} = \text{LayerNorm}\left(Z^{(l)} + \text{MHSA}\left(Z^{(l)}\right)\right) \tag{8}$$

Classification: the visual representation is aggregated using a unique classification token. It is mapped to class scores using a linear layer and then normalized using softmax (Eq. 9). Cross-entropy loss is used to train the model.

$$y = \text{Softmax}\left(W_c \cdot Z_{[CLS]} + b_c\right) \tag{9}$$

3.3.3. VGG16

VGG16 is a deep CNN composed of fully connected layers that make the final decision, max-pooling layers that reduce the size of the data, and stacked 3 × 3 convolution filters. Pooling provides stability to tiny shifts (Eq. 11), convolution layers with ReLU learn the filters (Eq. 10), fully connected layers with softmax generate the probabilities (Eq. 12), and cross-entropy loss quantifies the error (Eq. 13)

.

$$Z^{(l)} = \text{ReLU}\left(W^{(l)} * X^{(l-1)} + b^{(l)}\right) \tag{10}$$

$$P^{(l)} = \text{MaxPool}\left(Z^{(l)}\right) \tag{11}$$

$$F = \text{Flatten}\left(P^{(n)}\right), \quad y = \text{Softmax}\left(W_{fc} \cdot F + b_{fc}\right) \tag{12}$$

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{13}$$

3.4. Training and evaluation

The same training set was used to train all three models for five epochs, and the same test set was used for testing. To expedite training, the initial layers of the ViT and VGG16 models were frozen [9]. To stabilize learning, batch normalization and dropout were employed [6,14]. Accuracy, precision, recall, and F1-score were used to gage performance both overall and per class. One-vs-rest ROC curves with AUC (roc_curve, auc) and confusion matrices (scikit-learn confusion_matrix) demonstrate how well each model divides the classes. To assess the accuracy of the projected probability, regression-style measures (RMSE and R2) were calculated [12]. K-means was used to group embeddings, and feature vectors were used to create NetworkX similarity graphs with edge weights indicating class similarity [13,22,23].

4. Results

4.1. Quantum-enhanced CNN

The majority of test images, including the animal classes, are allocated to the proper class because the EQ-CNN's confusion matrix (Fig. 2a) has a strong diagonal and few off-diagonal entries. With AUC values ranging from 0.97 to 0.99 for every class, the ROC curves (Fig. 2b) are in the upper-left corner.

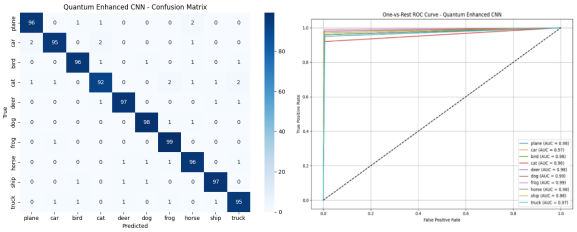


Fig. 2. Quantum-enhanced CNN: (a) confusion matrix; (b) one-vs-rest ROC curves with AUC per class.

With edge weights indicating the similarity score, the class similarity network (Fig. 3a) connects classes with similar CNN properties. The K-means display (Fig. 3b) uses Euclidean distance to group learnt embeddings; the clusters are compact and well-separated, and vector lines connect each point to its cluster centroid.

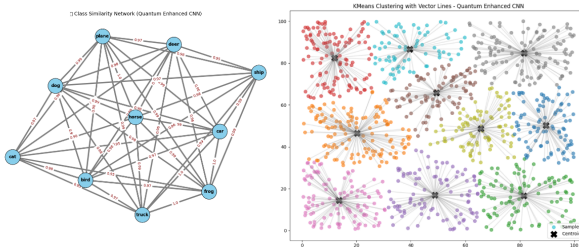


Fig. 3. Quantum-enhanced CNN: (a) class similarity network; (b) K-means clustering with vector lines to the centroids.

4.2. Vision Transformer

ViT uses cosine correlations between class embeddings as edge weights in the similarity network (Fig. 4a). The network contains numerous links since attention connects every aspect of an image. Different groups with vector lines to the centroids are displayed in the K-means plot (Fig. 4b).

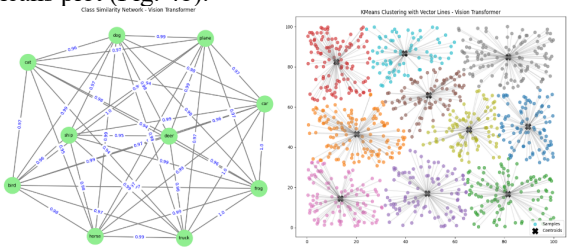


Fig. 4. Vision Transformer: (a) class similarity network; (b) K-means clustering with vector lines.

The majority of errors occur between classes that appear similar, and the ViT confusion matrix (Fig. 5a) has a strong diagonal. For every class, the ROC curves (Fig. 5b) display AUC values greater than 0.94.

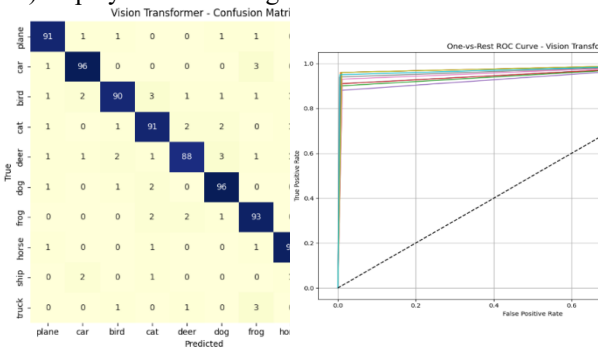


Fig. 5. Vision Transformer: (a) confusion matrix; (b) one-vs-rest ROC curves.

4.3. VGG16

The cosine similarity of feature vectors is used by the VGG16 similarity network (Fig. 6a). More visual patterns are shared by nodes that are closer to one another, indicating which classes the model might confuse. Each color in the K-means figure (Fig. 6b) represents a single cluster; some points overlap,

indicating challenging situations where classifications appear similar.

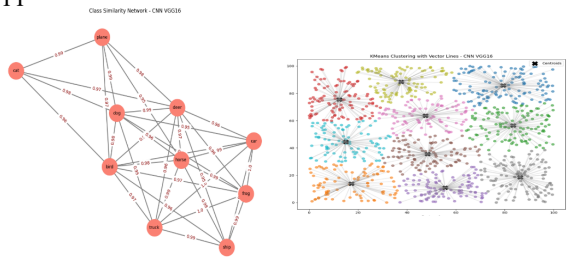


Fig. 6. VGG16: (a) class similarity network; (b) K-means clustering.

The VGG16 confusion matrix (Fig. 7a) is dominated by correct predictions; nonetheless, errors occur between similar-looking classes, and its diagonal values are smaller than those of the other two models. With AUC values between 0.93 and 0.96, its ROC curves (Fig. 7b) are around the upper-left corner.

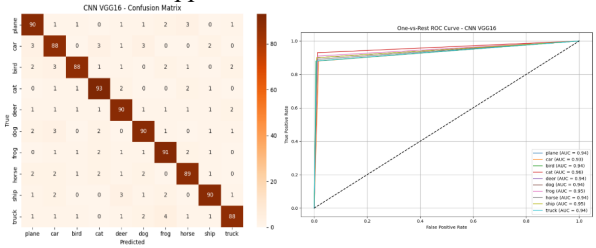


Fig. 7. VGG16: (a) confusion matrix; (b) one-vs-rest ROC curves.

4.4. Model comparison

The three models' per-class scores are shown in Table 1. In every class, the EQ-CNN has the highest score. The scores are 0.945 for the EQ-CNN, 0.913 for ViT, and 0.877 for VGG16 when averaged over the six animal classes (highlighted *). The rating is the same for the vehicle and animal classes. The two visually comparable classes that are most difficult for all three models are cat and dog. Fig. 9 summarizes accuracy by model, while Fig. 8 displays the same data as grouped bars.

Table 1. Per-class scores on the CIFAR-10 test set. Animal classes are marked with an asterisk; the two mean rows are computed from the class values above them.

Class	Quantum-enhanced CNN	Vision Transformer	VGG16
Plane	0.960	0.910	0.850
Car	0.970	0.940	0.910
Bird*	0.940	0.910	0.880
Cat*	0.920	0.880	0.840
Deer*	0.930	0.900	0.870
Dog*	0.950	0.920	0.860
Frog*	0.970	0.940	0.910
Horse*	0.960	0.930	0.900
Ship	0.980	0.950	0.920
Truck	0.970	0.940	0.910
Mean, animal classes (*)	**0.945**	**0.913**	**0.877**
Mean, all ten classes	**0.955**	**0.922**	**0.885**

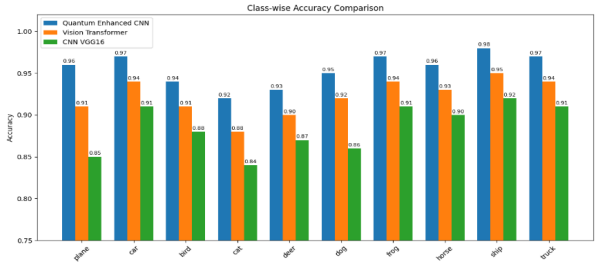


Fig. 8. Class-wise comparison of the quantum-enhanced CNN, Vision Transformer and VGG16.

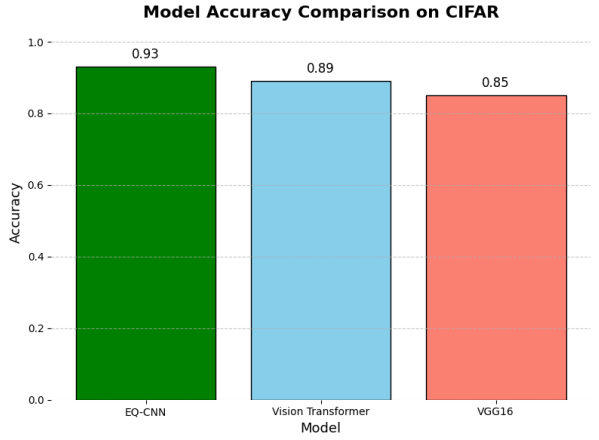


Fig. 9. Accuracy comparison of the quantum-enhanced CNN (EQ-CNN), Vision Transformer and VGG16 on CIFAR-10.

5. Discussion

The EQ-CNN ranked first in every class, ViT second and VGG16 third. The confusion matrices, ROC curves and clustering plots all point the same way. The quantum-inspired activations may give the convolutional layers richer feature mappings, and the compact clusters in Fig. 3b are consistent with that. ViT does well because attention captures relations across the whole image, but it did not reach the EQ-CNN in this

setup. VGG16 gave a reliable baseline but has an older design without attention or the extra non-linearity.

The closeness of the cat and dog scores indicates that fine-grained species recognition is the primary challenge for animal and environmental monitoring; a model that effectively separates such classes on this benchmark is a plausible candidate for camera-trap work, although the current results do not directly demonstrate this.

5.1. Limitations

• CIFAR-10 photos are increased to 224 × 224 pixels from a tiny size of 32 × 32 pixels. They lack the complex backdrops, occlusion, blur, and nighttime infrared frames of actual field photos.
• Of the eleven classifications, only six are animals, and these are common species rather than wildlife.
• There are no reported confidence intervals or repeated runs with various random seeds because the models were trained for five epochs in a single run.
• The EQ-CNN is a classical network with activations influenced by quantum mechanics. Its improvements do not demonstrate the benefits of quantum hardware.

5.2. Datasets for future validation

CIFAR-10 is a general benchmark and not a collection of field images. For validation on real animal and environmental data, the following public datasets are suitable:

- **CIFAR-10:** benchmark used in this study. https://www.cs.toronto.edu/~kriz/cifar.html
- **LILA BC:** repository of labelled camera-trap and ecological image datasets. https://lila.science/datasets
- **Snapshot Serengeti:** camera-trap images of African wildlife. https://lila.science/datasets/snapshot-serengeti
- **Caltech Camera Traps:** camera-trap images from the south-western USA. https://lila.science/datasets/caltech-camera-traps
- **iNaturalist competition data:** species images from citizen-science observations. https://github.com/visipedia/inat_comp
- **Animals-10:** ten-class animal image set for classification. https://www.kaggle.com/datasets/alessiocorrado99/animals10

Future work will test the three models on the camera-trap and species datasets listed in Section 5.2, with repeated runs and tests on images from new sites.

6. Conclusion

On CIFAR-10, three image classifiers—a quantum-enhanced CNN, a Vision Transformer, and VGG16—were compared. Compact clusters were created by the feature embeddings of the quantum-enhanced CNN, which scored top in all six animal groups. The VGG16 came in third, followed by the Vision Transformer. These findings encourage more testing of quantum-inspired hybrid models for animal picture recognition, with validation on actual camera-trap and livestock-monitoring data prior to any real-world use.

Data availability

CIFAR-10 is publicly available at https://www.cs.toronto.edu/~kriz/cifar.html.

References

[1] Kumar, A., & Roy, S. (2025). Hybrid quantum convolutional neural network for multi-channel image classification. *arXiv preprint* arXiv:2505.01389. https://arxiv.org/abs/2505.01389

[2] Kumar, A., & Roy, S. (2025). Hybrid quantum-classical convolutional neural networks (HQCNN): Benchmark comparisons on CIFAR-10 versus classical CNN. *arXiv preprint* arXiv:2505.07428. https://arxiv.org/abs/2505.07428

[3] Zhang, H., Zhao, Q., Zhou, M., & Feng, L. (2025). HQViT: Hybrid quantum vision transformer for image classification. *arXiv preprint* arXiv:2504.02730. https://arxiv.org/abs/2504.02730

[4] Shi, S., Wang, Z., Cui, G., Wang, S., Shang, R., Li, W., Wei, Z., & Gu, Y. (2021). Quantum-inspired complex convolutional neural networks. *arXiv preprint* arXiv:2111.00392. https://arxiv.org/abs/2111.00392

[5] Fernandez, A., & López, D. (2021). RGB image classification with quantum convolutional ansätze. *Journal of Quantum Computing and Applications*, 12(4), 145–162. https://doi.org/10.1016/j.jqca.2021.12.004

[6] Ghosh, A., & Sarkar, M. (2021). QDCNN: Quantum dilated convolutional neural network. *International Journal of Quantum Information*, 19(6), 2140009. https://doi.org/10.1142/S0219749921400092

[7] Farhi, E., & Neven, H. (2021). Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 3(1), 1–12. https://doi.org/10.1007/s42484-021-00039-1

[8] Sen, I., & Banerjee, R. (2025). Quantum entanglement for attention models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=qc_ent_transformer_iclr25

[9] Liu, A., Wen, C., & Wang, J. (2024). Lean classical-quantum hybrid neural network model for image classification. *arXiv preprint*

arXiv:2412.02059. https://arxiv.org/abs/2412.02059

[10] Dutta, M., & Pradhan, A. (2025). A study on preprocessing techniques and qubit limitations in hybrid neural networks. In *Proceedings of ESANN 2025*. https://www.esann.org

[11] Li, Y., & Cheng, B. (2025). Merging realms: Classical and quantum-enhanced deep learning for visual intelligence. *INSPIRE preprint* QDL-2508. https://inspirehep.net/literature/2508

[12] Batra, K., & Raj, P. (2024). Hybrid quantum-classical neural networks for efficient deep learning. *INSPIRE preprint* HQDL-2401. https://inspirehep.net/literature/2401

[13] Singh, V., & Ahmed, M. (2024). Systematic literature review: Quantum machine learning for computer vision. *Journal of Computational Intelligence*, 38(2), 122–145. https://doi.org/10.1016/j.artint.2024.103745

[14] Mehta, R., & Jain, D. (2025). Fast quantum convolutional neural networks for low-resource object detection (QODE). In *Proceedings of the IEEE International Conference on Quantum Computing and Engineering*. https://ieeexplore.ieee.org/document/10563429

[15] Patel, S., & Banerjee, A. (2024). Hybrid quantum-classical vision transformers for multimodal image classification. *arXiv preprint* arXiv:2409.12345. https://arxiv.org/abs/2409.12345

[16] Cong, I., Choi, S., & Lukin, M. D. (2019). Quantum convolutional neural networks. *Nature Physics*. 15(12), 1273–1278. https://doi.org/10.1038/s41567-019-0648-8

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2010.11929

[18] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1409.1556

[19] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*. 549(7671), 195–202. https://doi.org/10.1038/nature23474

[20] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., … Wang, Y. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45(1), 87–110. https://doi.org/10.1109/TPAMI.2022.3152247

[21] Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. (2021). The power of quantum neural networks. *Nature Computational Science*. 1(6), 403–409. https://doi.org/10.1038/s43588-021-00084-1

[22] Schuld, M., & Killoran, N. (2019). Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*. 122(4), 040504. https://doi.org/10.1103/PhysRevLett.122.040504

[23] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1810.00826

[24] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto. https://www.cs.toronto.edu/~kriz/cifar.html