

# Embedding-Aware Quantum-Classical SVMs for Scalable Quantum Machine Learning

Sebastián Andrés Cajas Ordóñez<sup>1,\*</sup>, Luis Fernando Torres Torres<sup>2</sup>, Mario Bifulco<sup>3</sup>,  
Carlos Andres Duran<sup>4</sup>, Cristian Bosch<sup>1</sup> and Ricardo Simon Carbajo<sup>1</sup>

<sup>1</sup>National Irish Centre for AI (CeADAR), University College Dublin (UCD), Dublin, Ireland

<sup>2</sup>SISTEMIC Research Group, University of Antioquia, Medellín, Colombia

<sup>3</sup>Department of Computer Science, University of Torino, Torino, Italy

<sup>4</sup>Corporation for Aerospace Initiatives (CASIRI), University of Cauca, Popayán, Colombia

## Abstract

Quantum Support Vector Machines face scalability challenges due to high-dimensional quantum states and hardware limitations. We propose an embedding-aware quantum-classical pipeline combining class-balanced *k-means* distillation with pretrained Vision Transformer embeddings. Our key finding: ViT embeddings uniquely enable quantum advantage, achieving up to 8.02% accuracy improvements over classical SVMs on Fashion-MNIST and 4.42% on MNIST, while CNN features show performance degradation. Using 16-qubit tensor network simulation via cuTensorNet, we provide the first systematic evidence that quantum kernel advantage depends critically on embedding choice, revealing fundamental synergy between transformer attention and quantum feature spaces. This provides a practical pathway for scalable quantum machine learning that leverages modern neural architectures.

## Keywords

Quantum Support Vector Machines (QSVMs), Hybrid Quantum-Classical Models, Pretrained Embeddings, Tensor Networks, Resource-Constrained Learning

## 1. Introduction

Quantum computing has emerged as a transformative paradigm with the potential to outperform classical approaches on specialized computational tasks [1]. Concurrently, machine learning (ML) continues advancing rapidly, driven by increasing data availability and accelerated computing hardware [2]. Quantum machine learning (QML), at the intersection of these fields, has significant potential to unlock new capabilities in data processing and complex problem solving [2, 3]. By utilizing quantum phenomena such as superposition and entanglement, QML algorithms may effectively address high dimensionality and combinatorial complexity beyond classical counterparts [4, 5].

Despite these prospects, large-scale quantum computing faces substantial practical challenges from noise and decoherence. Advanced error correction protocols can address these issues through sophisticated decoding algorithms, though their design complexity has led researchers to increasingly employ machine learning approaches for automation and enhancement of these decoding tasks [1]. Beyond error correction, QML applications have been explored extensively across domains such as image classification, natural language processing, and high energy physics [6, 7, 8, 9, 10]. Although proof of concept demonstrations on quantum processors like Google's Sycamore [11] and IBM's superconducting systems [12, 13] show feasibility, significant gaps persist between laboratory experiments and reliable industrial deployment [14, 15].

---

\*Corresponding author.

✉ sebastian.cajasordonez@ucd.ie (S. A. C. Ordóñez); lf.torres@udea.edu.co (L. F. T. Torres); mario.bifulco@unito.it (M. Bifulco); carlos.duran@unicauca.edu.co (C. A. Duran); cristian.boschserrano@ucd.ie (C. Bosch); ricardo.simoncarbajo@ucd.ie (R. S. Carbajo)

ORCID 0000-0003-0579-6178 (S. A. C. Ordóñez); 0009-0001-0706-0626 (L. F. T. Torres); 0009-0008-3243-7684 (C. A. Duran); 0000-0002-2121-2841 (R. S. Carbajo)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

A promising strategy for overcoming these gaps involves leveraging embedding and dimensionality reduction methods. Classical preprocessing techniques such as principal component analysis, or learned neural encoders like variational autoencoders, effectively reduce dataset complexity prior to quantum model input [8, 16]. These hybrid approaches help manage limited qubit resources on current quantum hardware while exploiting quantum enhanced feature spaces [17, 5]. Benchmark studies comparing quantum and classical ML approaches underline the potential and present limitations of QML [18, 14, 15]. Advancing this field requires systematic evaluations under realistic conditions, incorporating representative datasets, accurate noise models, and relevant performance metrics [2, 3, 19].

We propose an embedding-aware, hybrid quantum-classical QSVM framework designed to address the scalability limitations of quantum machine learning. By integrating class-balanced  $k$ -means data distillation with pretrained embeddings, our pipeline reduces data dimensionality while preserving task-relevant structure. Quantum kernel classification is performed using tensor network simulation with NVIDIA’s cuTensorNet [20, 21]. Benchmarking on MNIST and Fashion-MNIST shows that this embedding-driven approach consistently outperforms classical and quantum baselines in both accuracy and efficiency, confirming its value for scalable and resource-constrained quantum machine learning applications.

## **2. Related Work**

### **2.1. Scaling QML with Simulation Frameworks**

Simulating larger circuits remains a popular strategy because near-term quantum devices have limited qubits. Efficient tensor-network methods can push simulations of quantum support vector machines (QSVMs) to hundreds of qubits [21], addressing scaling issues that plague naïve state-vector simulators. This line of research proves instrumental for prototyping advanced QML algorithms, guiding their eventual deployment on real hardware [16, 22].

### **2.2. Quantum Classifiers for Image Recognition**

Variational quantum classifiers (VQCs), variational quantum circuits [6, 23, 22], quantum kernels, and hybrid architectures [7, 10, 3, 24, 25, 26] have been applied to standard benchmarks such as MNIST, Fashion-MNIST, and medical imaging tasks [6, 7, 9, 27]. Although classical deep networks often outperform small quantum models on large-scale datasets, quantum classifiers show competitive performance in data-scarce or high-dimensional settings by leveraging specialized embeddings and kernel methods [8, 26]. Several studies have also demonstrated end-to-end quantum classification pipelines executed on actual hardware, though typically limited to smaller datasets due to current device constraints [6, 10].

### **2.3. Hybrid Classical-Quantum Techniques**

Several researchers propose hybrid approaches: classical layers for data preprocessing or encoding, followed by quantum layers for feature transformation or classification [7, 8, 16]. These methods can offset limited qubit counts by handing only compressed or task-relevant information to the quantum circuit [6]. Genetic algorithms [16], autoencoders [8], and transfer learning [10] have all been employed to optimize these hybrid models.

### **2.4. Applications Beyond Image Classification**

QML has also been trialed in domains like high-energy physics [8, 13], medical diagnosis [7, 27], and scientific computing [21]. In certain specialized tasks - e.g., identifying the Higgs boson in proton collision data [8] - quantum models can match or exceed classical baselines under realistic noise models. These studies highlight the versatility of QML but also the pressing need for systematic benchmarking to compare cost-benefit trade-offs [18, 15].

As a whole, prior works illustrate QML’s broad applicability, from error decoding [1] and fundamental reviews [2, 14] to specialized classifiers for real-world tasks [6, 7, 10]. However, two critical gaps remain in the literature. First, a cohesive comparison that unifies insights across multiple domains and positions these results against robust classical baselines remains a key frontier [14, 15]. Second, existing approaches lack systematic investigation of how different embedding strategies affect quantum advantage, particularly the synergy between modern neural representations and quantum feature spaces.

The present study addresses both gaps by systematically evaluating quantum-enhanced classification models alongside classical baselines using diverse embedding strategies, revealing fundamental relationships between representation choice and quantum kernel performance on representative datasets.

### 3. Methodology

#### 3.1. Strategy Overview

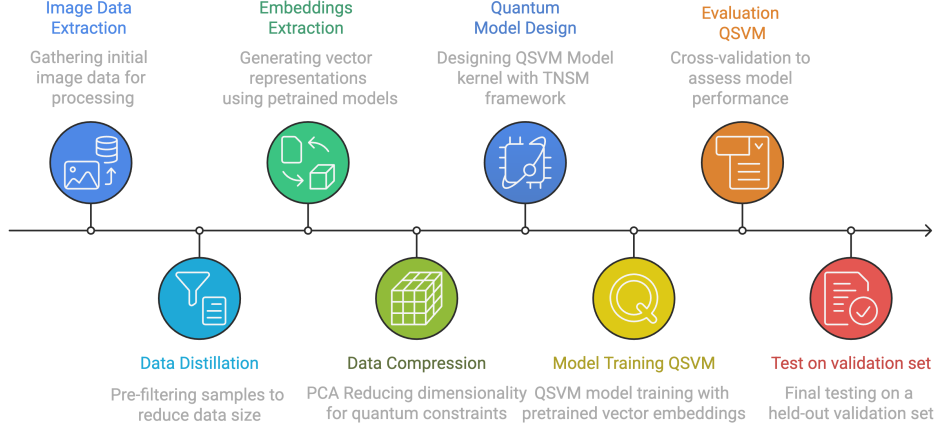
Our approach addresses the scalability challenges of quantum machine learning through a hybrid quantum-classical pipeline that strategically combines data preprocessing, feature extraction, and quantum kernel methods. As illustrated in Figure 1, the framework operates through eight sequential stages: we begin with image data extraction and preprocessing, followed by data distillation using class-balanced  $k$ -means clustering to reduce dataset size while maintaining representative samples. Next, we generate vector representations using pretrained models such as EfficientNet-B3 [28] and Vision Transformer variants [29], then apply Principal Component Analysis (PCA) to compress embeddings and match quantum hardware constraints. The processed embeddings are used to design a Quantum Support Vector Machine (QSVM) using the Tensor Network Support Matrix (TNSM) framework [21], which constructs quantum kernels through parameterized circuits and tensor network simulation using a data re-uploading and compute-uncompute strategy. Finally, we evaluate model performance through cross-validation and test on a held-out validation set to assess generalization capability. This embedding-aware strategy enables us to leverage the representational power of modern neural architectures while exploiting quantum kernel advantages for classification tasks, making quantum machine learning more practical and scalable for real-world applications.

#### 3.2. Quantum Model Architecture

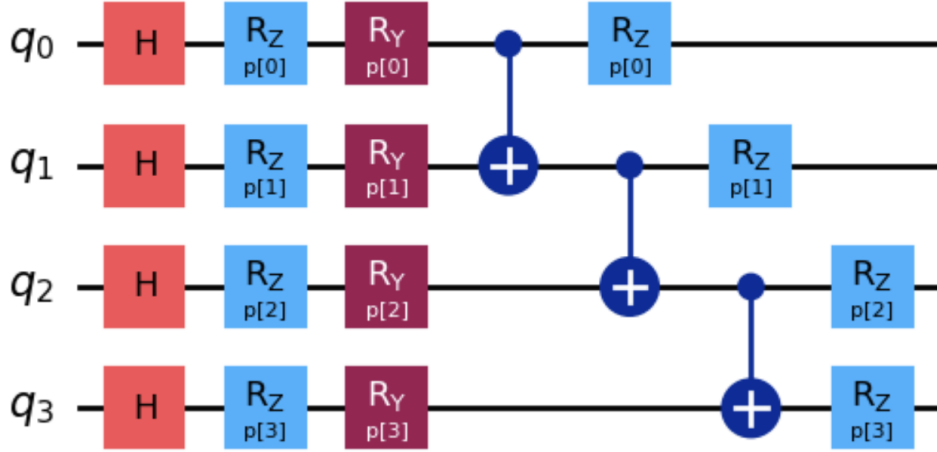
Chen et al. [21] forms the foundational simulation framework upon which our quantum model is built. The architecture employs a parameterized quantum circuit that encodes input data through rotational gates and entanglement layers within a Block-Encoded State (BPS) circuit, as shown in Figure 2. This data re-uploading circuit design has been validated in quantum learning applications and implemented within the Qiskit framework [30], serving as a proven foundation for kernel-based quantum classification.

It constructs a quantum kernel using a compute-uncompute strategy applied to the BPS circuit, parameterized on the input examples. This method enables the model to capture complex relationships between data points effectively. The resulting circuits are mapped onto tensor networks using the CircuitToEinsum converter, enabling efficient simulation on classical hardware. The kernel matrix is computed by contracting the tensor networks for all training pairs using an autotuned contraction path.

**Implementation Optimizations.** In contrast to the original implementation [21], we introduce several performance enhancements to the operand construction pipeline. First, to reduce redundant computations of trigonometric and exponential operations across repeated input angles, we apply function-level caching using Python’s @cache decorator. This memorization strategy significantly reduces overhead during the generation of gate matrices (e.g., parameterized  $Z$  and  $Y$  gates), which are frequently reused across multiple operand evaluations. Additionally, we precompute sine and cosine values in a shared utility function to avoid duplicating expressions and improve code modularity. Operand batches are generated via list comprehensions instead of iterative append calls, enhancing memory



**Figure 1:** Illustrates the sequential steps from data extraction to QSVN evaluation. The process begins with image data extraction, followed by class-balanced  $k$ -means clustering to distill representative samples. Vector embeddings are then extracted using ImageNet-pretrained models such as EfficientNet or ViT. To reduce dimensionality and match quantum hardware constraints, PCA is applied to compress the embeddings. These processed embeddings are used to design a Quantum Support Vector Machine (QSVN) using the TNSM framework, which constructs a quantum kernel via a data re-uploading and compute-uncompute strategy. The model is trained and validated through cross-validation, then evaluated on a held-out test set.



**Figure 2:** Quantum circuit used in the QSVN pipeline using Qiskit. Each of the four qubits is initialized with a Hadamard gate, followed by data-encoding rotations using parameterized  $R_Z$  and  $R_Y$  gates. A sequence of CNOT gates creates entanglement between adjacent qubits, after which a second layer of  $R_Z$  gates is applied. This structure forms an embedding-aware quantum feature map for encoding classical input features.

efficiency and readability. Likewise, tensor network amplitudes are computed using preallocated lists, eliminating dynamic resizing and reducing garbage collection pressure.

These optimizations form our enhanced baseline, designated as *Baseline+*, while the original implementation is referred to as *Baseline*. All comparisons throughout this paper use *Baseline+* as the reference. The improvements collectively reduce execution time and peak memory usage during simulation, as demonstrated in Table 3, and establish a foundation for future CuPy-based parallelization in operand template generation.

### 3.3. Quantum Kernel Formulation

The quantum kernel between two data points  $x_i$  and  $x_j$  is computed as the transition amplitude between their corresponding quantum states [3, 31]:

$$K_q(x_i, x_j) = |\langle \phi(x_i) | \phi(x_j) \rangle|^2 \quad (1)$$

where  $|\phi(x)\rangle = U(x)|0\rangle^{\otimes n}$  represents the quantum feature map implemented by our parameterized circuit  $U(x)$ . The quantum advantage emerges from the exponentially large Hilbert space dimension  $2^n$  compared to the classical feature space [4]. However, this advantage is critically dependent on how classical data  $x$  is embedded before quantum encoding.

## 4. Experimental Setup

### 4.1. Dataset and Data Distillation Process

Due to the computational complexity of scaling Quantum Support Vector Machines (QSVMs) to high-dimensional data, particularly in the context of image classification, we adopt a distilled version of the dataset to reduce resource requirements while preserving performance.

**MNIST Dataset:** We use the MNIST dataset [32], a widely recognized benchmark for evaluating image classification models. It consists of 70,000 grayscale images of handwritten digits (0–9), each of size  $28 \times 28$  pixels, divided into 60,000 training and 10,000 test samples.

**Fashion-MNIST Dataset:** We also utilize the Fashion-MNIST dataset [33], a benchmark dataset designed as a more challenging alternative to MNIST. It comprises 70,000 grayscale images of 10 fashion item categories (e.g., t-shirts, trousers, dresses), each of size  $28 \times 28$  pixels, split into 60,000 training and 10,000 test samples. This dataset provides a diverse set of visual patterns, enabling robust evaluation of our QSVM pipeline in a multi-class classification setting.

**Data Distillation:** To address QSVM scalability constraints, we employ a class-balanced dataset distillation approach based on  $k$ -means clustering. The algorithm iterates through each class, applies  $k$ -means with  $k = 200$  to identify representative centroids, and selects the real data point closest to each centroid as a prototype, yielding exactly 200 samples per class. The resulting distilled dataset contains 2,000 samples total (1,600 for training, 400 for testing), reducing computational complexity from  $\mathcal{O}(70000^2)$  to  $\mathcal{O}(1600^2)$  kernel evaluations while preserving representative coverage of each class’s feature distribution and eliminating class imbalance effects. The distillation parameters ( $k$  value and dataset size) can be customized in our implementation based on available computational resources and hardware constraints, enabling adaptation to different quantum simulation capabilities.

### 4.2. Embedding Extraction and Data Compression

To construct compact and informative inputs for quantum classification, we extract high-dimensional feature embeddings using pretrained image encoders. Specifically, we employ EfficientNet-B3 [28] and Vision Transformer (ViT) variants [29] trained via the CLIP framework [34]. These models, pretrained on large-scale datasets, capture rich semantic features that are well-suited for downstream classification tasks.

EfficientNet-B3 produces 1536-dimensional embeddings, while ViT models typically output 768 or 512-dimensional vectors. To evaluate trade-offs between representation richness and simulation cost, we experiment with three dimensionality settings: 512, 768, and 1536, across different architectures. For lower-dimensional settings, we apply Principal Component Analysis (PCA) to reduce embedding size while preserving key variance, thereby aligning inputs with the capacity limits of our 16-qubit quantum kernel simulator.

To benchmark performance across embedding strategies, we map each encoder-dimension pair to a shorthand label used throughout our analysis (e.g., ViT-B/16-512, EffNet-1536). Table 1 summarizes these configurations, including the native embedding size and dimensionality used in simulation.



As a point of reference, we include two baselines: the `Baseline`, based on Chen et al.’s original QSVM using flattened image pixels, and our enhanced version, `Baseline+`, which incorporates the computational enhancements introduced in Section 3.2. All remaining models use the same enhanced QSVM backend as `Baseline+`, differing only in their embedding source and size.

**Table 1**

Model categories and configurations used in the experiments. `Baseline` refers to the original implementation by Chen et al. [21], using flattened pixel inputs. `Baseline+` (ours) is our enhanced QSVM implementation using raw pixels. All other variants apply the same enhanced QSVM pipeline with different pretrained embeddings.

Label	Description	Input Dim.
<code>Baseline</code>	Chen et al. with flattened pixels	784
<code>Baseline+</code> (ours)	Enhanced QSVM pipeline with raw pixels	784
QSVM: <code>EffNet-512</code>	EfficientNet-B3 embedding	512
QSVM: <code>EffNet-1536</code>	EfficientNet-B3 embedding	1536
QSVM: <code>ViT-B/32-512</code>	ViT-B/32 (CLIP) embedding	512
QSVM: <code>ViT-B/16-512</code>	ViT-B/16 (CLIP) embedding	512
QSVM: <code>ViT-L/14</code>	ViT-L/14 (CLIP) embedding	768
QSVM: <code>ViT-L/14@336</code>	ViT-L/14@336px (CLIP) embedding	768

### 4.3. Evaluation Methodology

Model performance is assessed through 5-fold cross-validation to ensure robust statistical evaluation. We measure classification accuracy, precision, F1-score, and Area Under the Curve (AUC) to provide comprehensive performance characterization. Computational efficiency is evaluated by tracking total execution time and peak memory usage during training and evaluation phases.

Classical SVMs are implemented using scikit-learn’s SVC with RBF kernel and hyperparameters ( $C=1.0$ ,  $\gamma=\text{'scale'}$ ,  $\text{probability}=\text{True}$ ). All preprocessing steps, including embedding extraction and PCA dimensionality reduction, are identical between classical and quantum approaches. This ensures that performance differences reflect only the kernel computation method rather than data preparation artifacts.

Our evaluation directly contrasts quantum support vector machines against classical SVM baselines using identical feature representations and evaluation protocols. This approach isolates the impact of quantum kernel methods from preprocessing effects, enabling fair assessment of quantum advantage claims. Statistical significance is evaluated through cross-validation consistency and standard deviation analysis.

### 4.4. Computational Infrastructure

All experiments are conducted on NVIDIA A100 Tensor Core GPUs with 80GB HBM2 memory, using CUDA 12.0 and NVIDIA’s cuQuantum cuTensorNet backend [20] for quantum simulation. This high-performance computing environment ensures consistent benchmarking conditions and enables efficient tensor network contraction for quantum kernel computation. The GPU-accelerated simulation framework allows us to explore larger quantum circuits than would be feasible with CPU-based approaches.

## 5. Results and Analysis

### 5.1. Quantum Advantage with Modern Neural Embeddings

Our central finding demonstrates that quantum support vector machines achieve consistent performance improvements over classical SVMs when using Vision Transformer embeddings, while showing degraded

performance with raw pixels or CNN vector embeddings for this specific setting. Table 2 presents our key results: Quantum models with ViT embeddings achieve accuracy gains up to 4.4% on MNIST and 8.0% on Fashion-MNIST, while traditional approaches (raw pixels, EffNet features) show performance degradation.

*Baseline+* reduced runtime from 4,492 to 3,812 seconds during cross-validation, saving 680 seconds, and brought peak memory usage down from 44.1GB to 43.5GB. All enhancements to the quantum pipeline, including caching, memory preallocation, and parallel tensor contractions, were built on this enhanced baseline rather than the original, as observed in Table 3.

**Table 2**

Quantum vs Classical SVM Performance Comparison. Held-out test accuracy demonstrating quantum advantage with modern neural embeddings. Quantum advantage represents the relative improvement.

	Dataset	Embedding type	Classic SVM Acc	Baseline+ (ours) Acc	Quantum Advantage
MNIST		Raw Pixels	<b>0.945</b>	0.887	-6.14
		EffNet-512	<b>0.9694</b>	0.935	-3.55
		EffNet-1536	<b>0.9731</b>	0.948	-2.58
		ViT-B/32-512	0.9481	<b>0.990</b>	<b>+4.42</b>
		ViT-B/16-512	0.9544	<b>0.995</b>	<b>+4.25</b>
		ViT-L/14	0.9825	<b>0.990</b>	<b>+0.76</b>
		ViT-L/14@336-768	0.9838	<b>0.993</b>	<b>+0.94</b>
FMNIST		Raw Pixels	<b>0.7825</b>	0.730	-6.71
		EffNet-512	<b>0.9172</b>	0.887	-3.29
		EffNet-1536	<b>0.916</b>	0.877	-4.26
		ViT-B/32-512	0.8476	<b>0.900</b>	<b>+6.18</b>
		ViT-B/16-512	0.8332	<b>0.900</b>	<b>+8.02</b>
		ViT-L/14	0.8708	<b>0.897</b>	<b>+3.01</b>
		ViT-L/14@336-768	0.8652	<b>0.900</b>	<b>+4.02</b>

This quantum advantage emerges specifically with transformer-based representations, revealing a fundamental synergy between quantum kernels and modern neural embeddings. All Vision Transformer variants demonstrate positive quantum advantage, indicating this is not merely an artifact of specific architectural choices but rather reflects a deeper compatibility between quantum feature spaces and transformer-learned representations.

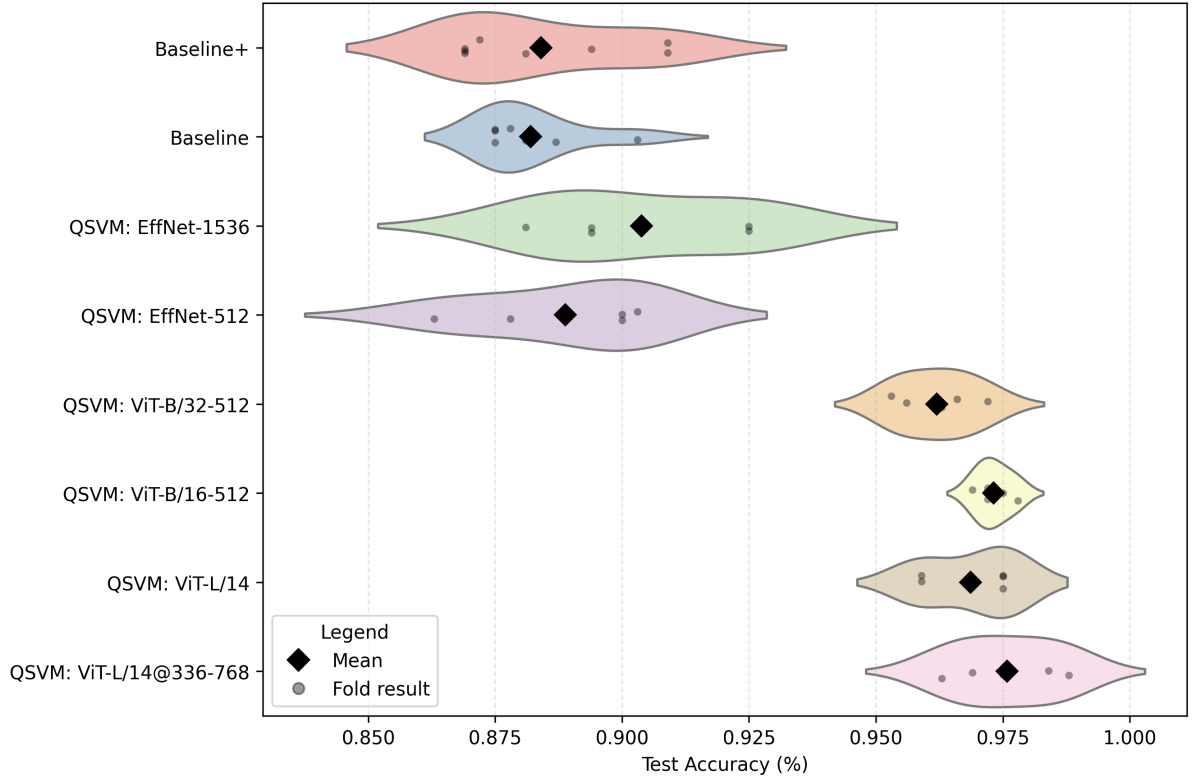
The results highlight the critical importance of feature representation selection in quantum machine learning. Traditional approaches using raw pixels or CNN-based features consistently favor classical methods, while transformer embeddings unlock quantum computational advantages. This finding has important implications for the design of quantum machine learning systems, suggesting that the preprocessing stage is as crucial as the quantum algorithm itself for achieving quantum advantage.

The practical significance of up to 8% accuracy improvement represents substantial value for real-world applications, particularly in domains where high precision is critical such as medical diagnosis or safety-critical systems. These gains, while seemingly modest, can translate to significant improvements in deployment scenarios where accuracy differences directly impact outcomes.

## 5.2. Cross-Validation Performance Analysis

Comprehensive 5-fold cross-validation results confirm the robustness of our quantum advantage findings across multiple performance metrics. Table 3 presents detailed performance evaluation including accuracy, precision, F1-score, AUC, runtime, and memory usage for both classical baselines and quantum-enhanced models.

The best-performing quantum model, QSVM using ViT-L/14@336-768, achieves an accuracy of 97.6% on MNIST and 84.1% on Fashion-MNIST, clearly surpassing the baselines over pixel information, which level off around 88.2% and 72.5%, respectively. The near-perfect AUC scores of 99.9% across all ViT-based



**Figure 3:** Violin plots show the distribution of test accuracy across K-fold cross-validation for MNIST. The width of each violin indicates the density of results; wider sections reflect more frequent accuracy values, helping visualize consistency and variability in model performance.

**Table 3**

Benchmark Results for MNIST and FashionMNIST using 16 qubits. Highest values for accuracy-based metrics and lowest for runtime and memory usage are in **bold**.

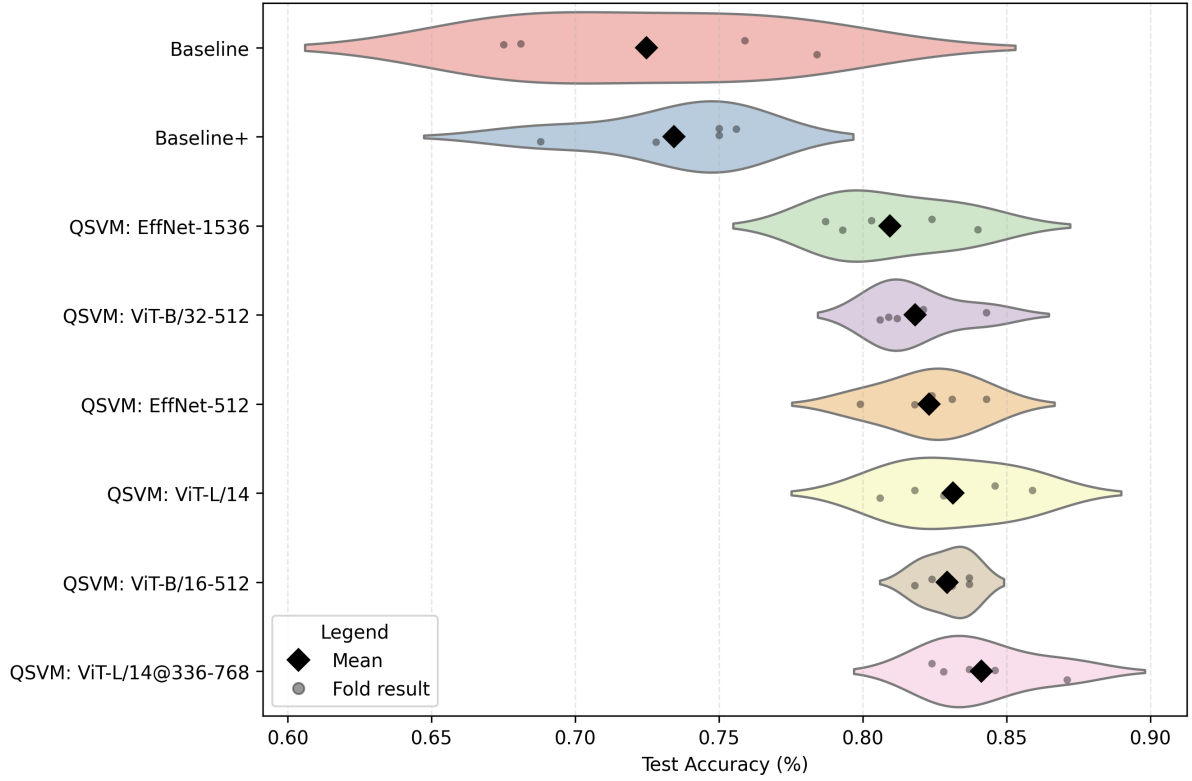
Dataset	Model	Test Acc	Precision	F1	AUC	Time (s)	Memory (MB)
MNIST	Baseline	$0.882 \pm 0.010$	$0.887 \pm 0.010$	$0.882 \pm 0.011$	$0.990 \pm 0.004$	$4492.196 \pm 39.285$	$44116.842 \pm 25.978$
	Baseline+	$0.884 \pm 0.018$	$0.888 \pm 0.019$	$0.884 \pm 0.018$	$0.991 \pm 0.004$	$3812.316 \pm 42.187$	$43537.845 \pm 22.515$
	QSVM: EffNet-512	$0.889 \pm 0.018$	$0.893 \pm 0.015$	$0.889 \pm 0.017$	$0.992 \pm 0.003$	$3910.851 \pm 25.007$	<b><math>43506.193 \pm 21.365</math></b>
	QSVM: EffNet-1536	$0.904 \pm 0.020$	$0.906 \pm 0.019$	$0.904 \pm 0.020$	$0.994 \pm 0.003$	$3819.504 \pm 23.488$	$43566.972 \pm 22.614$
	QSVM: ViT-B/32-512	$0.962 \pm 0.008$	$0.963 \pm 0.007$	$0.962 \pm 0.008$	$0.999 \pm 0.000$	$3900.742 \pm 24.954$	$43510.314 \pm 21.536$
	QSVM: ViT-B/16-512	$0.973 \pm 0.003$	$0.974 \pm 0.003$	$0.973 \pm 0.003$	$0.999 \pm 0.000$	<b><math>3763.170 \pm 25.646</math></b>	$43513.467 \pm 20.800$
	QSVM: ViT-L/14	$0.969 \pm 0.009$	$0.970 \pm 0.008$	$0.969 \pm 0.008$	<b><math>0.999 \pm 0.001</math></b>	$3816.003 \pm 31.957$	$43520.979 \pm 18.243$
	<b>QSVM: ViT-L/14@336-768</b>	<b><math>0.976 \pm 0.010</math></b>	<b><math>0.977 \pm 0.010</math></b>	<b><math>0.975 \pm 0.010</math></b>	<b><math>0.999 \pm 0.001</math></b>	$3939.404 \pm 24.480$	$43520.375 \pm 22.726$
FashionMNIST	Baseline	$0.725 \pm 0.048$	$0.723 \pm 0.041$	$0.716 \pm 0.044$	$0.963 \pm 0.003$	$4456.288 \pm 32.991$	$44086.054 \pm 22.615$
	Baseline+	$0.734 \pm 0.028$	$0.727 \pm 0.029$	$0.723 \pm 0.027$	$0.963 \pm 0.004$	$3803.786 \pm 27.142$	$43510.356 \pm 19.410$
	QSVM: EffNet-512	$0.823 \pm 0.016$	$0.823 \pm 0.019$	$0.818 \pm 0.016$	$0.980 \pm 0.002$	$3797.365 \pm 29.575$	$43256.111 \pm 21.782$
	QSVM: EffNet-1536	$0.809 \pm 0.022$	$0.808 \pm 0.020$	$0.805 \pm 0.020$	$0.980 \pm 0.004$	$3887.396 \pm 26.549$	$43301.836 \pm 17.939$
	QSVM: ViT-B/32-512	$0.818 \pm 0.015$	$0.821 \pm 0.014$	$0.816 \pm 0.015$	$0.981 \pm 0.002$	$3773.245 \pm 25.367$	<b><math>43250.348 \pm 24.488</math></b>
	QSVM: ViT-B/16-512	$0.829 \pm 0.008$	$0.831 \pm 0.009$	$0.827 \pm 0.009$	$0.982 \pm 0.004$	$3853.586 \pm 38.180$	$43258.243 \pm 23.672$
	QSVM: ViT-L/14	$0.831 \pm 0.021$	$0.831 \pm 0.022$	$0.829 \pm 0.022$	$0.981 \pm 0.003$	<b><math>3766.821 \pm 21.742</math></b>	$43266.337 \pm 20.614$
	<b>QSVM: ViT-L/14@336-768</b>	<b><math>0.841 \pm 0.019</math></b>	<b><math>0.841 \pm 0.020</math></b>	<b><math>0.840 \pm 0.020</math></b>	<b><math>0.983 \pm 0.002</math></b>	$3859.313 \pm 20.656$	$43265.254 \pm 20.394$

quantum models suggest that these approaches reliably capture discriminative patterns with minimal classification errors.

The consistent improvement across cross-validation folds is especially notable. Accuracy standard deviations remain low, typically between  $\pm 0.003$  and  $\pm 0.020$ , showing that the observed quantum advantage is stable and reproducible, rather than the result of favorable data splits or initialization. This level of consistency is vital for deployment in practical settings where reliable performance is required.

In addition, the quantum models display strong precision-recall alignment, with precision scores





**Figure 4:** Violin plots show the distribution of test accuracy across K-fold cross-validation for FashionMNIST. The width of each violin indicates the density of results; wider sections reflect more frequent accuracy values, helping visualize consistency and variability in model performance.

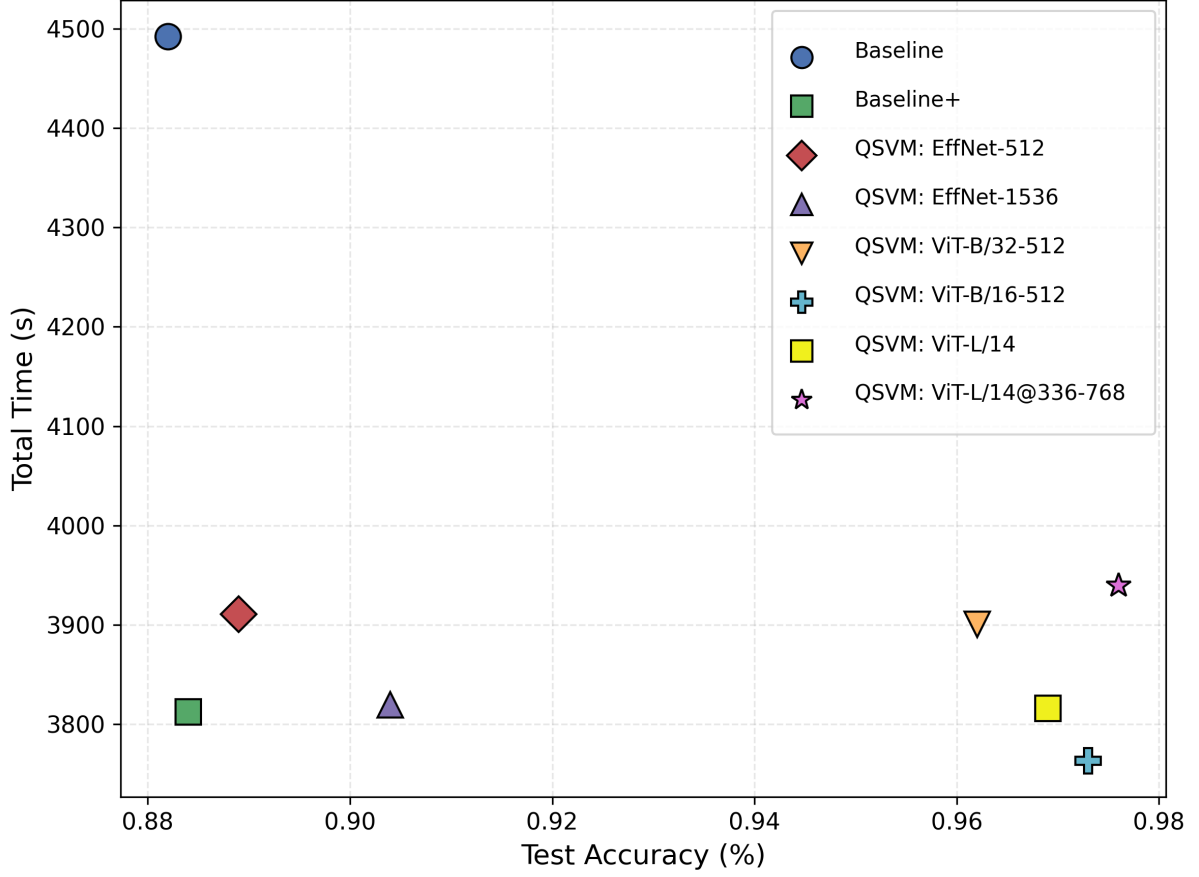
closely tracking overall accuracy across all configurations. This balanced performance suggests that quantum kernels provide meaningful gains in class-level discrimination, rather than boosting accuracy by favoring specific categories.

The confusion matrices in Figures 7 and 8 further illustrate the generalization power of our top model, QSVM with ViT-L/14@336-768, showing alignment between cross-validation and held-out test results. Generalization is stronger for MNIST than for Fashion-MNIST. The clear diagonal structure and few off-diagonal errors confirm that the high accuracy reflects true performance across all digit classes, not just select ones.

Violin plots in Figures 3 and 4 visualize test accuracy distributions over cross-validation folds. QSVM models with ViT embeddings, including ViT-B/16, ViT-L/14, and ViT-L/14@336, consistently achieve higher average accuracies and lower variance compared to both baselines and EfficientNet-based QSVMs. ViT-B/16-512 and ViT-L/14@336-768 show especially narrow, high-accuracy distributions on MNIST, while baseline and EfficientNet models display wider, more variable spreads, particularly on Fashion-MNIST. These results highlight the advantage of transformer embeddings in delivering stable and accurate quantum classification across tasks of varying complexity.

### 5.3. Computational Efficiency and Scalability Analysis

Our embedding-enhanced quantum models demonstrate strong accuracy while maintaining reasonable computational demands for quantum simulations. Most ViT-based quantum configurations complete training and evaluation in approximately 3,800 seconds with consistent memory usage around 43GB, as detailed in Table 3. While these runtimes may appear substantial, they represent a significant improvement over prior quantum simulations and are reasonable given the high-dimensional embedding spaces and tensor contraction overhead inherent to quantum kernel methods.



**Figure 5:** Comparison of total execution time and test accuracy for different QSVM models for MNIST. The x-axis represents the average test accuracy across K-folds, while the y-axis (log scale) shows the total runtime in seconds. Each point corresponds to a model variant, with horizontal and vertical lines indicating the standard deviation of accuracy and time, respectively.

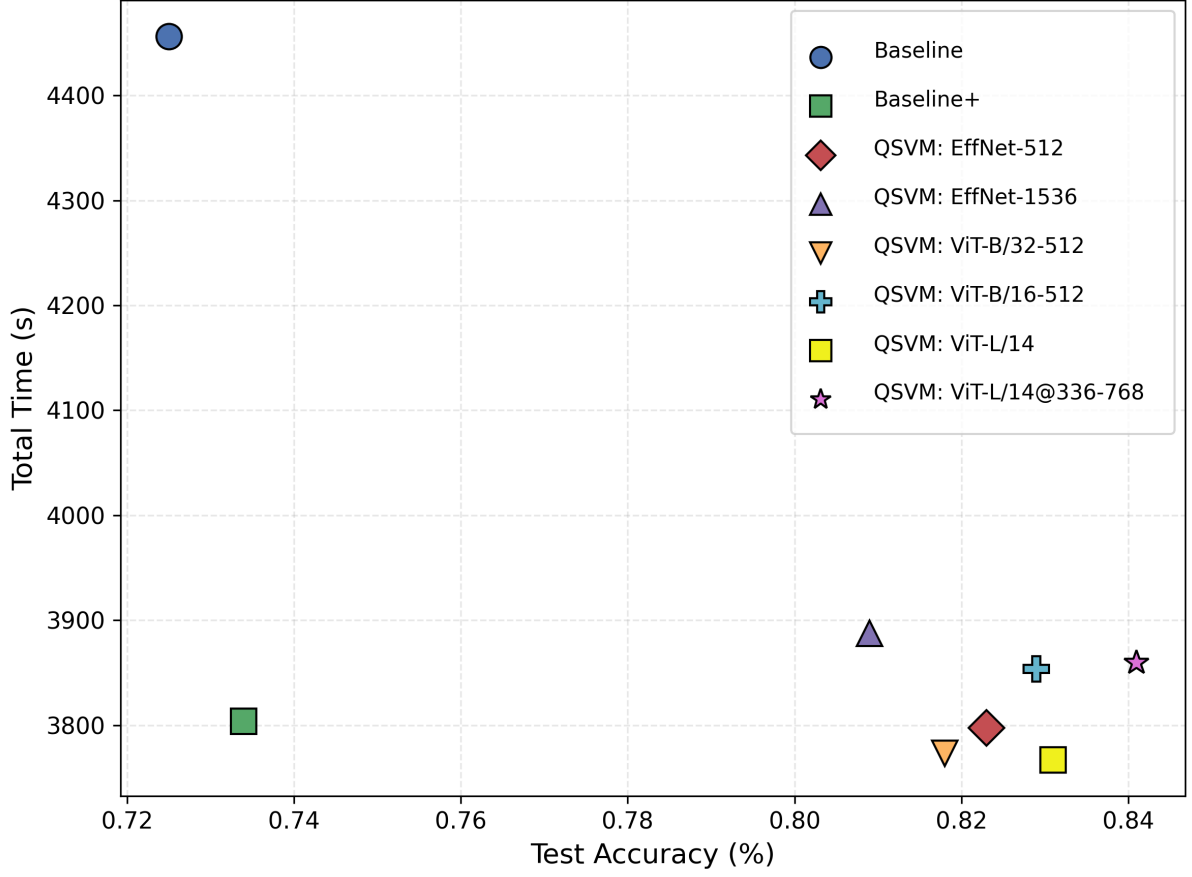
Among the top-performing models, QSVM with ViT-B/16-512 offers the optimal balance between performance and efficiency, achieving 97.3% accuracy with the fastest runtime of 3,763 seconds. Figures 5 and 6 illustrate the trade-offs between computational cost and classification performance. Vision Transformer-based models consistently achieve top-tier accuracy with moderate computational requirements, while EfficientNet configurations provide competitive accuracy with reduced resource demands. The original *Baseline* model shows the least favorable performance-efficiency balance, while *Baseline+* demonstrates consistent runtime improvements.

These results confirm that embedding-enhanced quantum models offer practical scalability alongside significant accuracy gains. Vision Transformer embeddings clearly outperform EfficientNet-B3 across both datasets, and the computational overhead of quantum kernel methods is effectively mitigated by the reduced dataset sizes achieved through strategic data distillation, making these approaches viable for real-world deployment.

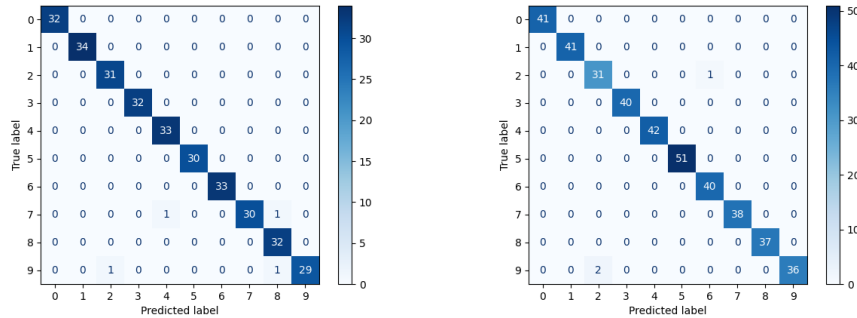
## 6. Discussion

This study demonstrates that quantum advantage in machine learning emerges not from quantum algorithms alone, but from the strategic synergy between quantum kernels and appropriate feature representations. Our central finding reveals that Vision Transformer embeddings uniquely unlock quantum advantage, achieving up to 8.02% accuracy improvements over classical SVMs, while CNN features and raw pixels consistently favor classical approaches.

Although quantum simulations demand substantial computational resources (approximately 3,800



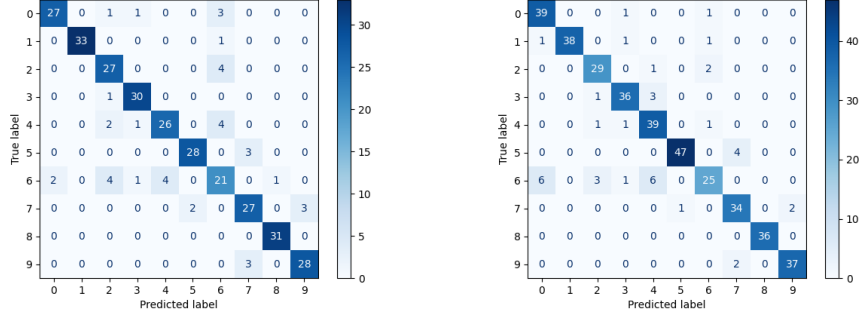
**Figure 6:** Comparison of total execution time and test accuracy for different QSVM models for FMNIST. The x-axis represents the average test accuracy across K-folds, while the y-axis (log scale) shows the total runtime in seconds. Each point corresponds to a model variant, with horizontal and vertical lines indicating the standard deviation of accuracy and time, respectively.



(a) Validation Fold (Best CV Model) for MNIST dataset (b) Held-out Test Set (Best CV Model) for MNIST dataset

**Figure 7:** (a) Performance on the validation fold used to select the best model. (b) Performance of that model on the held-out test set. This comparison highlights how well the selected model generalizes to unseen data.

seconds for training), this investment proves justified in high-precision applications where accuracy improvements directly translate to enhanced outcomes, particularly in medical diagnosis, safety-critical systems, and fraud detection scenarios. The computational overhead becomes advantageous compared to scaling classical approaches for similar accuracy gains, which typically require exponentially larger datasets or increasingly complex architectures. Our strategic data distillation effectively reduces problem



(a) Validation Fold (Best CV Model) for Fashion-MNIST dataset (b) Held-out Test Set (Best CV Model) for Fashion-MNIST dataset

**Figure 8:** (a) Performance on the validation fold used to select the best model. (b) Performance of that model on the held-out test set. This comparison highlights how well the selected model generalizes to unseen data.

complexity from  $\mathcal{O}(70000^2)$  to  $\mathcal{O}(1600^2)$  kernel evaluations, making quantum kernel methods tractable while preserving essential dataset characteristics.

Our framework demonstrates that quantum machine learning achieves scalability through intelligent preprocessing, with distillation parameters easily customizable based on available computational resources. This adaptability enables flexible deployment across diverse scenarios, from resource-constrained environments utilizing smaller distilled datasets to high-performance settings leveraging full quantum simulation capabilities. The embedding-aware approach establishes a practical pathway toward quantum advantage that becomes increasingly favorable as quantum hardware continues to mature.

## 7. Conclusion

We present an embedding-aware quantum-classical framework that systematically addresses scalability challenges in quantum machine learning by strategically combining class-balanced data distillation with pretrained embeddings. Building upon Chen et al.’s GPU-accelerated quantum kernel method [21], our pipeline successfully reduces computational complexity while achieving measurable performance improvements over classical baselines.

Our work delivers the first systematic evidence that quantum kernel advantage depends critically on embedding choice, revealing fundamental compatibility between transformer attention mechanisms and quantum feature spaces. Through 16-qubit tensor network simulation, we demonstrate consistent quantum advantages using ViT embeddings across MNIST (up to 4.42% improvement) and Fashion-MNIST (up to 8.02% improvement), while observing performance degradation with CNN-based features.

The embedding-aware QSVM framework enables practical quantum machine learning deployment through configurable data distillation and hardware-adaptive preprocessing. Compared to raw inputs, structured transformer embeddings consistently deliver superior accuracy and generalization, effectively supporting real-world applications in high-dimensional classification tasks where precision remains critical.

**Limitations and Future Directions.** Several limitations require attention for broader impact. Our evaluation concentrates on relatively simple visual classification benchmarks (MNIST, Fashion-MNIST); validation on complex datasets such as CIFAR-10, medical imaging, or domain-specific applications remains necessary to assess generalization. The theoretical foundations explaining transformer-quantum synergy remain largely underexplored, presenting compelling opportunities for fundamental research.

Future work should pursue automated embedding and kernel selection strategies [35] to eliminate manual hyperparameter tuning, explore sophisticated dimensionality reduction techniques beyond PCA to better preserve semantic information, and develop optimized quantum circuit designs [36]

for enhanced computational efficiency. Expanding empirical validation to medical imaging and other high-dimensional domains will prove critical for demonstrating broader practical utility.

This work establishes that achieving quantum advantage requires careful algorithm-representation co-design rather than naive application of quantum methods. Our embedding-aware framework provides both immediate practical value for precision-critical applications and a scalable foundation for quantum machine learning that effectively leverages modern neural architectures. As quantum hardware continues to mature, this approach offers a viable pathway toward practical quantum advantage in real-world machine learning applications.

## 8. Data and Code Availability

The code used in this study is publicly available at: <https://github.com/sebasmos/QuantumVE>.

## 9. Acknowledgments

This work was supported by the Google Cloud Research Credits program under the award number GCP19980904.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

- [1] J. Bausch, A. W. Senior, F. J. Heras, T. Edlich, A. Davies, M. Newman, C. Jones, K. Satzinger, M. Y. Niu, S. Blackwell, et al., Learning high-accuracy error decoding for quantum processors, *Nature* (2024) 1–7.
- [2] D. Peral-García, J. Cruz-Benito, F. J. García-Peñalvo, Systematic literature review: Quantum machine learning and its applications, *Computer Science Review* 51 (2024) 100619.
- [3] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, *Nature* 567 (2019) 209–212.
- [4] A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, S. Woerner, The power of quantum neural networks, *Nature Computational Science* 1 (2021) 403–409.
- [5] A. Senokosov, A. Sedykh, A. Saginalieva, B. Kyriacou, A. Melnikov, Quantum machine learning for image classification, *Machine Learning: Science and Technology* 5 (2024) 015040.
- [6] K. Shen, B. Jobst, E. Shishenina, F. Pollmann, Classification of the fashion-mnist dataset on a quantum computer, *arXiv preprint arXiv:2403.02405* (2024).
- [7] A. K. K. Don, I. Khalil, M. Atiquzzaman, A fusion of supervised contrastive learning and variational quantum classifiers, *IEEE Transactions on Consumer Electronics* 70 (2024) 770–779.
- [8] V. Belis, P. Odagiu, M. Grossi, F. Reiter, G. Dissertori, S. Vallecorsa, Guided quantum compression for high dimensional data classification, *Machine Learning: Science and Technology* 5 (2024) 035010.
- [9] X. Vasques, H. Paik, L. Cif, Application of quantum machine learning using quantum kernel algorithms on multiclass neuron m-type classification, *Scientific Reports* 13 (2023) 11541.
- [10] J. Kim, J. Huh, D. K. Park, Classical-to-quantum convolutional neural network transfer learning, *Neurocomputing* 555 (2023) 126643.
- [11] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, et al., Quantum advantage in learning from experiments, *Science* 376 (2022) 1182–1186.
- [12] G. Gentinetta, A. Thomsen, D. Sutter, S. Woerner, The complexity of quantum support vector machines, *Quantum* 8 (2024) 1225.

- [13] D. Cugini, D. Gerace, P. Govoni, A. Perego, D. Valsecchi, Comparing quantum and classical machine learning for vector boson scattering background reduction at the large hadron collider, *Quantum Machine Intelligence* 5 (2023) 35.
- [14] Y. Gujju, A. Matsuo, R. Raymond, Quantum machine learning on near-term quantum devices: Current state of supervised and unsupervised techniques for real-world applications, *Physical Review Applied* 21 (2024) 067001.
- [15] D. Basilewitsch, J. F. Bravo, C. Tutschku, F. Struckmeier, Quantum neural networks in practice: A comparative study with classical models from standard data sets to industrial images, *arXiv preprint arXiv:2411.19276* (2024).
- [16] K. Phalak, A. Ghosh, S. Ghosh, Optimizing quantum embedding using genetic algorithm for qml applications, *arXiv preprint arXiv:2412.00286* (2024).
- [17] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, D. Scheiermann, R. Wolf, Training deep quantum neural networks, *Nature communications* 11 (2020) 808.
- [18] C. Havenstein, D. Thomas, S. Chandrasekaran, Comparisons of performance between quantum and classical machine learning, *SMU Data Science Review* 1 (2018) 11.
- [19] R. Potempa, S. Porebski, Comparing concepts of quantum and classical neural network models for image classification task, in: *Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021* 12, Springer, 2022, pp. 61–71.
- [20] NVIDIA Corporation, NVIDIA cuTensorNet: High-Performance Tensor Network Library, 2024. <https://docs.nvidia.com/cuda/cuquantum/latest/cutensornet/index.html>.
- [21] K.-C. Chen, T.-Y. Li, Y.-Y. Wang, S. See, C.-C. Wang, R. Wille, N.-Y. Chen, A.-C. Yang, C.-Y. Lin, Validating large-scale quantum machine learning: Efficient simulation of quantum support vector machines using tensor networks, *Machine Learning: Science and Technology* (2024).
- [22] K. Mitarai, M. Negoro, M. Kitagawa, K. Fujii, Quantum circuit learning, *Physical Review A* 98 (2018) 032309.
- [23] E. Farhi, H. Neven, Classification with quantum neural networks on near term processors, *arXiv preprint arXiv:1802.06002* (2018).
- [24] D. Sharma, P. Singh, A. Kumar, The role of entanglement for enhancing the efficiency of quantum kernels towards classification, *Physica A: Statistical Mechanics and its Applications* 625 (2023) 128938.
- [25] B.-S. Chen, J.-L. Chern, Generating quantum feature maps for svm classifier, *arXiv preprint arXiv:2207.11449* (2022).
- [26] C. Blank, D. K. Park, J.-K. K. Rhee, F. Petruccione, Quantum classifier with tailored quantum kernel, *npj Quantum Information* 6 (2020) 41.
- [27] W. E. Maouaki, T. Said, M. Bennai, Quantum support vector machine for prostate cancer detection: A performance analysis, *arXiv preprint arXiv:2403.07856* (2024).
- [28] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020).
- [30] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, et al., Quantum computing with qiskit, *arXiv preprint arXiv:2405.08810* (2024).
- [31] M. Schuld, N. Killoran, Quantum machine learning in feature hilbert spaces, *Physical review letters* 122 (2019) 040504.
- [32] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* 86 (1998) 2278–2324.
- [33] H. Xiao, K. Rasul, R. Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, *arXiv preprint arXiv:1708.07747* (2017).
- [34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin,



- J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.
- [35] M. Incudini, D. L. Bosco, F. Martini, M. Grossi, G. Serra, A. D. Pierro, Automatic and effective discovery of quantum kernels, *IEEE Transactions on Emerging Topics in Computational Intelligence* (2024) 1–10. URL: <http://dx.doi.org/10.1109/TETCI.2024.3499993>. doi:10.1109/tetci.2024.3499993.
- [36] L. Sünkel, D. Martyniuk, D. Mattern, J. Jung, A. Paschke, Ga4qco: Genetic algorithm for quantum circuit optimization, 2023. URL: <https://arxiv.org/abs/2302.01303>. arXiv:2302.01303.