Quantum Probabilistic Label Refining: Enhancing Label Quality for Robust Image Classification

Fang Qi
Computer Science
Tulane University
New Orleans, LA, USA
fqi2@tulane.edu

Lu Peng
Computer Science
Tulane University
New Orleans, USA
lpeng3@tulane.edu

Zhengming Ding
Tulane University
NEW ORLEANS, LA, USA
zding1@tulane.edu

Abstract

Learning with softmax cross-entropy on one-hot labels often leads to overconfident predictions and poor robustness under noise or perturbations. Label smoothing mitigates this by redistributing some confidence uniformly, but treats all samples equally, ignoring intra-class variability. We propose a hybrid quantum–classical framework that leverages quantum non-determinism to refine data labels into probabilistic ones, offering more nuanced, human-like uncertainty representations than label smoothing or Bayesian approaches. A variational quantum circuit (VQC) encodes inputs into multi-qubit quantum states, using entanglement and superposition to capture subtle feature correlations. Measurement via the Born rule extracts probabilistic soft labels that reflect input-specific uncertainty. These labels are then used to train a classical convolutional neural network (CNN) with soft-target cross-entropy loss. On MNIST and Fashion-MNIST, our method improves robustness—achieving up to 50% higher accuracy under noise—while maintaining competitive clean-data accuracy. It also enhances model calibration and interpretability, as CNN outputs better reflect quantum-derived uncertainty. This work introduces **Quantum Probabilistic Label Refining**, which bridges quantum measurement and classical deep learning to enable robust training via refined, correlation-aware labels, without architectural changes or adversarial techniques.

CCS Concepts

• **Computer systems organization** → *Quantum computing*; • **Computing methodologies** → Machine learning.

Keywords

Quantum computing, probabilistic labels, label refinement

ACM Reference Format:
Fang Qi, Lu Peng, and Zhengming Ding. 2026. Quantum Probabilistic Label Refining: Enhancing Label Quality for Robust Image Classification. In *Great Lakes Symposium on VLSI 2026 (GLSVLSI '26), June 22–24, 2026, Canandaigua, NY, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3787109.3815246

GLSVLSI '26, Canandaigua, NY, USA

ACM ISBN 979-8-4007-2431-2/26/06
https://doi.org/10.1145/3787109.3815246

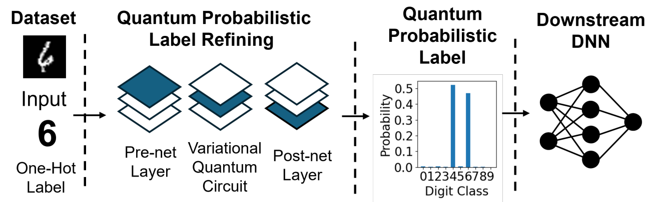


Figure 1: Conceptual overview of our QPLR approach, where an input digit is encoded into a variational quantum circuit and outputs a *probabilistic label distribution* over the $K = 10$ classes. This soft label supervises a downstream DNN to capture more realistic crowd-source knowledge.

1 Introduction

Deep learning has greatly improved image classification, yet even top models remain brittle under small distribution shifts such as noise, blur, or rotation [18, 26]. In contrast, human perception remains robust and well-calibrated under similar conditions [12, 20, 26, 28]. One reason is that models are typically trained with one-hot labels, which treat classification as deterministic and enforce full confidence in a single class, even for ambiguous inputs [33]. This overconfidence is risky in high-stakes domains: in autonomous driving, misclassifying a partially occluded road sign can be dangerous, while in medical imaging, confident errors on borderline cases may lead to misdiagnosis. Recognizing and communicating uncertainty is therefore essential [13, 19].

Label smoothing is a common solution that redistributes a fraction of confidence from the true class to others to reduce overconfidence [32]. However, it applies the same softened distribution to all samples, ignoring differences in ambiguity among inputs. In reality, class overlaps and uncertain samples are poorly represented in standard datasets. A data-dependent method to reflect sample-specific uncertainty—something static label smoothing cannot achieve.

To address this, we turn to quantum non-determinism as a natural source of uncertainty. Quantum systems yield probabilistic outputs upon measurement—a principle formalized by the Born rule [14], which assigns outcome probabilities based on the squared amplitude of quantum states [29]. By encoding images into a variational quantum circuit (VQC), we harness quantum superposition and entanglement to produce rich, sample-specific distributions over class labels [10, 24]. Specifically, entanglement encodes input-feature correlations, and superposition represents diverse configurations. Quantum measurement then extracts input-dependent probabilistic labels that reflect these correlations, offering a more expressive alternative to classical heuristics.

We introduce Quantum Probabilistic Label Refining (QPLR), a novel framework that uses a VQC to generate soft labels from quantum measurement distributions, shown in Fig. 1. These are then used to train a classical convolutional neural network (CNN) via soft-target cross-entropy loss. Unlike one-hot or uniform-smoothed labels, these targets reflect input-specific ambiguity and class overlap. To our knowledge, this is the first work to leverage quantum measurement distributions as soft labels for classical deep learning. Our contributions are summarized as follows:

- We propose QPLR, a hybrid method that uses quantum measurements from variational quantum circuits (VQCs) to produce refined, sample-specific probabilistic labels.
- We show that CNNs trained with QPLR achieve up to 50% higher accuracy under image corruptions on MNIST and Fashion-MNIST without sacrificing clean-data performance.
- We demonstrate that QPLR improves uncertainty estimation and interpretability for safety-critical applications such as autonomous driving and medical AI.

2 Related Work

Quantum Deep Learning. Variational Quantum Circuits (VQCs) have emerged as a novel approach for realizing near-term quantum advantage in supervised learning tasks [5, 8, 23]. Compared to traditional quantum neural networks (QNNs) [1, 10, 17] that rely on mid-circuit measurements or deep architectures, VQCs prioritize resource efficiency on NISQ devices by using shallow parameterized circuits without intermediate state collapse, thereby mitigating decoherence and measurement noise [3, 8]. Early works successfully demonstrated VQCs on small-scale benchmarks such as MNIST, achieving feasibility on simulators and noisy hardware [7, 9, 15, 25, 31, 34]. However, these efforts have focused on using the quantum model itself as the final classifier—i.e., collapsing the measurement outcome into a single label or expectation value. In contrast, our work exploits the *full probabilistic output* of the VQC (i.e., the distribution over measurement outcomes) as a *teaching signal* for a classical network.

Quantum–Classical Hybrid Learning. Several recent efforts have integrated quantum circuits into classical neural pipelines, for instance, as trainable feature maps [1, 31] or hybrid layers [16, 17]. These hybrid models typically incorporate quantum subroutines directly into the forward pass of the classifier [2, 6, 11, 30]. In contrast, our hybridization occurs at the *label level*: we treat the quantum circuit as a *black-box label generator*. Once soft labels are generated—potentially offline on quantum hardware or high-performance simulators—the main training remains entirely classical and scalable on future quantum devices. To our knowledge, using quantum measurement distributions to supervise classical deep networks and exploit Born-rule randomness for improved learning is novel.

3 Motivation

While prior hybrid quantum–classical work has explored variational quantum circuits (VQCs) for classification, few have investigated their potential for *label refinement*. We argue that quantum mechanisms—**entanglement**, **superposition**, and **measurement**—make VQCs uniquely suited to generate soft, probabilistic labels that reflect ambiguity and inter-class relationships in real-world data.

Entanglement allows quantum circuits to encode rich correlations between qubits representing different image regions or features. For n qubits, there are $2^n - n - 1$ possible non-trivial entangled subsets. At $n = 10$, this yields 1013 distinct correlation structures, including pairwise, triplet, and global entanglement. Such expressiveness enables modeling of complex dependencies that classical models struggle to represent.

Superposition, achieved through quantum gates (e.g., R_Y, R_Z, U_3), encodes classical inputs into a 2^n-dimensional Hilbert space. This permits simultaneous representation of all basis states. Combined with entanglement, it enables highly expressive quantum states, allowing the circuit to explore varied label interpretations and capture fine-grained intra-class variation.

Measurement, governed by the Born rule, collapses the quantum state into a class probability distribution. Unlike one-hot labels or fixed label-smoothing heuristics, these measurement-induced distributions vary by input and reflect the uncertainty encoded by the quantum state. They provide a principled way to produce *probabilistic labels* that convey richer, sample-wise supervision.

Recent theoretical works validate this advantage. Huang et al. [21] demonstrate that random quantum circuits can produce kernel functions beyond the reach of classical approximations. Abbas et al. [1] show that quantum models can linearly separate classes in Hilbert spaces that are otherwise inaccessible to classical networks. These insights suggest that quantum circuits offer powerful data representations with exponentially fewer parameters.

4 Proposed Method

4.1 Quantum Probabilistic Label Refining

Our first step is to train a *Variational Quantum Circuit* (VQC) on the MNIST training set and then use it as a *quantum labeler*. Concretely, let $x \in \mathbb{R}^{28\times28}$ be an input digit reshaped to a vector in $\mathbb{R}^d$ (with $d = 28^2$). We allocate n qubits so that $2^n \geq K = 10$, and define two possible encodings:

Angle encoding: A classical feed-forward block maps x to an n-dimensional rotation vector $\boldsymbol{\phi}(x) \in [0, \pi]^n$. We then prepare

$$|\psi(x)\rangle \;=\; \bigotimes_{i=1}^{n} R_Y\big(\phi_i(x)\big)\;|0\rangle^{\otimes n},$$

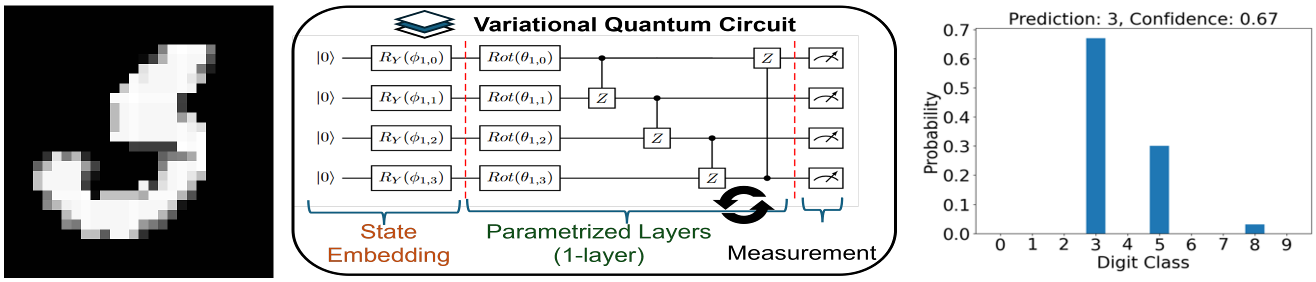
where $R_Y(\phi)$ is a Pauli-Y rotation.

Amplitude encoding: We first flatten and normalize x to a 2^n-dimensional state vector $v(x)$, then embed via $|\psi(x)\rangle \;=\; \sum_{i=0}^{2^n-1} v_i(x)\;|i\rangle$.

On $|\psi(x)\rangle$ we apply a *parameterized circuit* of L layers, each consisting of

$$\underbrace{\bigotimes_{i=1}^{n} \mathrm{Rot}(\theta_{l,i})}_{\text{single-qubit rotations}} \qquad \text{followed by} \qquad \underbrace{\prod_{(i,j)\in\mathcal{E}} \mathrm{CZ}_{i,j}}_{\text{controlled-Z entanglers}},$$

where the graph $\mathcal{E}$ encodes either a *linear*, *ring*, or *full* entanglement pattern. The resulting state $|\psi(x;\boldsymbol{\theta})\rangle \;=\; U(\boldsymbol{\theta})\,V(x)\;|0\rangle^{\otimes n}$ depends on both the input-encoding unitary $V(x)$ and the trainable unitary $U(\boldsymbol{\theta})$. Finally, we perform M repeated projective measurements in



(a) (b) (c)

Figure 2: An illustrative example of generating a soft Quantum labeling. (a) An ambiguous MNIST digit "5". (b) A simple 4-qubit variational quantum circuit composed of state embedding, one layer of parametrized circuit, and measurement. (c) The soft label contains distributions that provide both 3 and 5 as the predictions with higher probability than other digits.

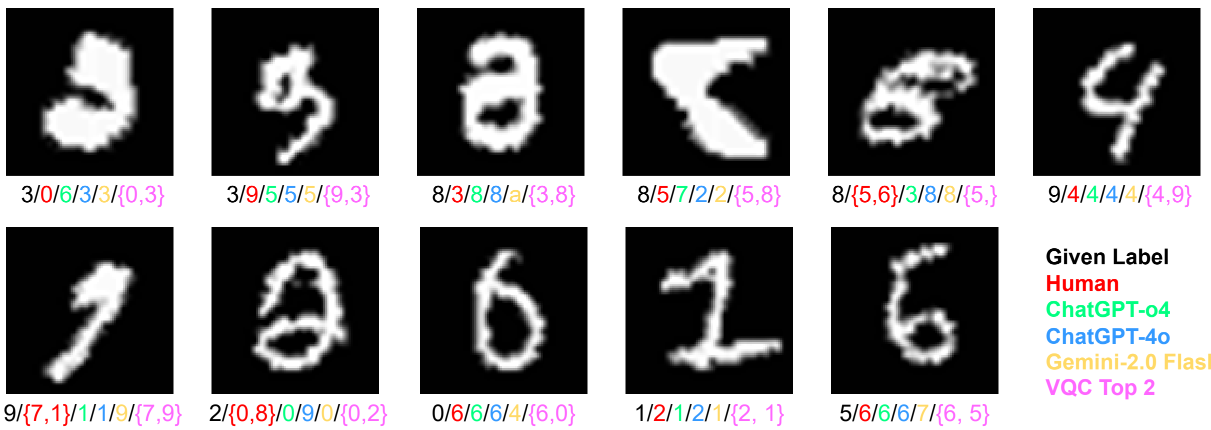


Figure 3: Comparison of human, foundation AI model, and quantum probabilistic predictions on ambiguous MNIST train examples. For each digit image, we report the digit recognition assigned by human annotators, ChatGPT-4, ChatGPT-4o, and Gemini 2.0 Flash, and the top 2 classes within the VQC's distribution. These results illustrate differing perceptions of uncertainty across models and humans. [Best viewed in color.]

the computational basis $\{|y\rangle\}_{y=0}^{2^n-1}$, yielding outcome y with Born-rule probability $P(y \mid x) = \left|\langle y|\ \psi(x;\boldsymbol{\theta})\right|^2$.

We then truncate or renormalize these probabilities to the $K = 10$ digit classes. In practice, we sample $M = 1{,}000$ shots with PennyLane's simulator (or hardware) to estimate the distribution $\{P(y \mid x)\}$ without any manual calibration.

Because quantum superposition allows $|\psi(x)\rangle$ to overlap multiple basis states, the VQC naturally produces *non-deterministic* labels. For a confusing training example (e.g. one labeled "5" that visually resembles a "3"), the estimated distribution might be $P(5 \mid x) \approx 0.67, \quad P(3 \mid x) \approx 0.31, \quad P(y \neq 3, 5) \approx 0.02$, reflecting genuine ambiguity, with an example shown as Fig. 2. In contrast, a classical softmax-based CNN forced to one-hot targets would likely assign $\approx 99\%$ to a single class. Moreover, the smooth manifold of quantum states in Hilbert space ensures that small perturbations of x yield gradual changes in $|\psi(x)\rangle$, so similar inputs have similar label distributions.

We train the VQC's parameters $\boldsymbol{\theta}$ via gradient-based optimization (adjoint differentiation on GPU, parameter-shift otherwise) to minimize the cross-entropy

$$\mathcal{L}_{\mathrm{Q}}(\boldsymbol{\theta}) = -\sum_{i=1}^{N}\sum_{k=0}^{K-1} y_k^{(i)} \log P(k \mid x^{(i)}; \boldsymbol{\theta}),$$

where $y_k^{(i)} \in \{0, 1\}$ are the one-hot true labels. Although the VQC's classification accuracy ($\sim 96\%$) may lag that of a CNN, our goal is *not* to beat classical performance but to extract a *probabilistic* supervision signal for downstream training.

4.2 Human and Foundation AI Assessment

To evaluate the interpretability and reliability of our quantum-derived soft labels, we conducted a two-pronged assessment involving both expert human annotators and foundation models. We selected ambiguous MNIST samples from the training dataset whose quantum-generated label distributions exhibited high entropy (see Figure 3).

Human Assessment. The authors independently reviewed each grayscale digit image and assessed its identity based on visual inspection. Using a majority voting scheme, the most likely digit class was selected. In instances where an image exhibited visual ambiguity—appearing similar to more than one digit—multiple plausible classes were recorded. The first digit listed represents the class deemed most likely, while secondary options reflect residual uncertainty. This approach provides a human-centric benchmark for evaluating label ambiguity and comparing against both quantum-derived distributions and AI model predictions.

Foundation Model Assessment. We queried three state-of-the-art AI models ChatGPT-4, ChatGPT-4o, and Gemini-2.0 Flash—using the prompt: *"Please recognize the image as one of the 10 handwritten digit classes."*

As shown in Fig.3, for genuinely ambiguous inputs, the ground-truth label and conventional model outputs sometimes diverge from human perception, whereas our quantum-derived distributions faithfully capture that uncertainty. The VQC assigns its highest probability to the class that humans favor, while still allocating a nonzero mass to the dataset's annotated label. By contrast, foundation models like Gemini occasionally ignore prompt constraints—for instance, predicting a non-digit category "a" for an image labeled "2"—likely reflecting internal pattern-recognition biases. In another case, an image annotated as "8" but judged by experts as a "5" or "6" was assigned > 99.9% probability on "5" by the VQC, illustrating its close alignment with human judgment even when it conflicts with the original annotation. These observations demonstrate that even state-of-the-art AI systems can struggle with edge-case ambiguity, highlighting the benefit of probabilistic labeling grounded in quantum nondeterminism.

4.3 Deep Image Classification

In image classification tasks, a deep neural network (DNN) typically begins by receiving an input image $\mathbf{x}_i \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the image height, width, and number of channels, respectively. A neural network backbone, such as a convolutional neural network (CNN), processes this input to extract high-level feature representations. These features are then passed to a classifier head—often a fully connected layer followed by a softmax function—to produce a probability distribution over K classes $\mathbf{p}_i = \mathcal{F}_\theta(\mathbf{x}_i) \in [0,1]^K$, where $\mathcal{F}_\theta$ denotes the DNN with parameters θ, and $\mathbf{p}_i$ represents the predicted class probabilities for the i-th sample. The classifier is typically trained using the cross-entropy (CE) loss on one-hot labeled data. Let $\mathbf{y}_i \in \{0,1\}^K$ be the one-hot ground-truth label for the i-th input. The CE loss is defined as:

$$\mathcal{L}_{\mathrm{ce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij}\log(p_{ij}) = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{ik}), \tag{1}$$

where p_{ik} is the prediction for the correct class k of sample i.

However, standard supervised learning typically assumes access to deterministic, one-hot labels $\mathbf{y}_{\text{one-hot}}$, which encode full confidence in a single class. Training models with softmax cross-entropy loss on such labels often results in overconfident predictions and reduced robustness to input noise or distributional shifts.

The standard softmax cross-entropy loss penalizes deviations from the single "correct" class and ignores the ambiguity inherent in inputs that may resemble multiple classes (e.g., a digit that could reasonably be a "3" or an "8"), meaning one-hot targets can lead to models that are overly confident, often resulting in poor generalization. To mitigate this, *label smoothing* [32] replaces the hard 0/1 targets with smoothed labels, assigning a small portion of probability mass to incorrect classes. This technique acts as a form of regularization, reducing overfitting and encouraging the model to produce more calibrated probability estimates.

4.3.1 Probabilistic Labeling via Quantum Non-Determinism. In real-world classification scenarios, a single input x may correspond to multiple plausible outcomes $y \in \{1,\ldots,K\}$. To better reflect this uncertainty, we propose using *probabilistic labels* $\mathbf{y}^{\text{prob}} \in \Delta^{K-1}$, where each label is a probability distribution over the class space. While label smoothing uniformly adjusts all labels by redistributing a fixed amount of confidence, it fails to account for sample-specific ambiguity and treats all inputs identically.

To address this, we introduce a principled mechanism to generate probabilistic labels using **quantum non-determinism**. Specifically, we employ a variational quantum circuit (VQC) to encode each input x into a quantum state $|\psi(x)\rangle$, and extract a class distribution by measuring the state and applying the Born rule [14]. This produces a quantum-derived soft label: $\mathbf{y}^{\text{quantum}} = [p_1^{\text{q}},\ldots,p_K^{\text{q}}]$, where each p_k^{q} represents the probability of observing class k given the quantum encoding of x.

We then train a classical neural network using the following soft-label cross-entropy loss $\mathcal{L}'_{\mathrm{ce}} = -\sum_{k=1}^{K} y_k^{\text{quantum}} \log p_k$, where $\mathbf{p}$ is the predicted distribution from the model. This approach encourages the network to distribute its confidence across plausible classes, leading to improved robustness under noise, better calibration, and more human-like uncertainty estimation.

Such probabilistic modeling is particularly valuable in high-stakes applications. In autonomous driving, visual inputs may be ambiguous due to weather, occlusion, or sensor noise, and committing to a single class (e.g., traffic sign type or pedestrian intent) can be dangerous. A probabilistic prediction enables safer decision-making by allowing the system to hedge its actions based on uncertainty.

5 Experiments

5.1 Experimental Setting

5.1.1 Datasets. We evaluate our methods on two widely used image classification benchmarks.

The MNIST dataset contains 70,000 grayscale images of handwritten digits (0–9), each sized 28×28 pixels. These images are centered and size-normalized, with 60,000 for training and 10,000 for testing. Fashion-MNIST shares the same format and train/test split as MNIST but features 10 categories of clothing items (e.g., shirts, trousers, sneakers). The dataset is significantly more challenging due to its greater visual complexity and intra-class variation, offering a more realistic benchmark for evaluating generalization to natural image textures and shapes.

To better assess model robustness and generalization, we create two challenging variants of these datasets: (a) **Noisy Variants**: We apply additive Gaussian noise with zero mean and a standard deviation to simulate sensor imperfections or real-world degradation. (b) **Rotated Variants**: We rotate each image by a fixed angle to evaluate the model's resilience to geometric transformations that often occur in practical settings.

5.1.2 Implementation Details.

Quantum Implementation. Our variational quantum circuits (VQCs) are defined and executed in PennyLane v0.25 [4], with a lightweight PyTorch–based pre- and post-network wrapped around each quantum kernel. All experiments ran on the LONI HPC, where each

Table 1: Test accuracy of different label refining methods (including BNN and RS) under increasing Gaussian noise.

Method	Std=0.1	Std=0.2	Std=0.3	Std=0.4	Std=0.5
M1	99.14%	98.12%	85.50%	62.59%	37.68%
M2	99.30%	98.60%	73.51%	51.01%	23.32%
M3	97.87%	97.35%	95.76%	81.34%	70.13%
M4	97.95%	97.82%	96.05%	80.66%	52.82%
BNN	98.72%	95.31%	75.04%	45.55%	46.82%
RS	98.82%	87.02%	66.98%	58.59%	37.63%

node provides two NVIDIA A100 GPUs and 32 CPU cores. For both MNIST and Fashion-MNIST, each image $\mathbf{x}_i \in \mathbb{R}^{28\times 28}$ is flattened to a 784-dimensional vector and passed through a two-layer pre-network (128 units per layer, ReLU activations) that produces $n = 10$ rotation angles for *angle encoding* into a 10-qubit register. On MNIST, we stack $L = 3$ variational layers—each comprising single-qubit R_Y rotations followed by ring entanglement via Controlled Z gates—and on Fashion-MNIST we use $L = 2$. Circuits are measured with $M = 1{,}000$ shots, and the resulting counts are renormalized by a two-layer post-network (128⟶10) to yield a probability distribution over the ten different classes. Parameters $\boldsymbol{\theta}$ are trained with Adam (initial Learning Rate as 0.001, decayed ten-fold after epoch 3) for five epochs on MNIST (batch size 64), reaching 97% accuracy in approximately 13 hours, and for 25 epochs on Fashion-MNIST, achieving 88.93% accuracy in around 54 hours.

Networks Implementation. We implement a LeNet-style CNN for digit classification on MNIST and its variants. The model consists of two convolutional layers: the first applies 6 filters of size 5×5, followed by ReLU activation and 2×2 max pooling, while the second applies 16 filters of size 5×5, again followed by ReLU and 2×2 max pooling. The resulting feature maps are flattened and passed through two fully connected layers with 120 and 84 units, respectively, before reaching the final output layer with 10 units, corresponding to the digit classes. The model is trained using the Adam optimizer with a learning rate of 0.0001 and standard cross-entropy loss, optionally with label smoothing. All experiments are conducted on a machine equipped with two NVIDIA Ada 6000 GPUs, and end-to-end trainings are completed in roughly 10 minutes. We fixed random seeds for NumPy, PyTorch, and PennyLane to ensure deterministic behavior across runs. Code and hyperparameter details are available in our supplementary repository.

5.2 Comparison Results

Model Variants. We evaluate several variants of a LeNet-style CNN on MNIST and Fashion-MNIST. **M1** is the baseline model trained with standard hard labels and Softmax classification. **M2** extends M1 with label smoothing [27], which regularizes training by slightly softening the target labels. **M3** replaces one-hot targets with probabilistic labels generated by our Quantum Probabilistic Label Refinement (QPLR) using all training samples. **M4** is a variant of M3 that uses only high-confidence samples (confidence > 0.9, 54,887 of 60k) to reduce noise during training.

Additional Baselines. Beyond label smoothing, we compare against two uncertainty-based label refinement methods. **Bayesian Neural Networks (BNN)** use MC Dropout to generate soft labels by averaging predictions across stochastic forward passes, capturing epistemic uncertainty. **Randomized Smoothing (RS)** generates soft labels by averaging predictions over multiple Gaussian-noised inputs ($\sigma = 0.05$), capturing local input-space uncertainty. The resulting soft labels are used to train downstream CNNs, evaluated under noise and rotation corruptions . In summary, **M1** = hard labels, **M2** = label smoothing, **M3/M4** = QPLR variants, and **BNN/RS** are uncertainty-based baselines.

Table 2: Label refining methods test performance under combined input noise and rotation (20°).

Method	Std= 0.1+20°	Std= 0.2+20°	Std= 0.3+20°	Std= 0.4+20°	Std= 0.5+20°
M1	95.24%	91.58%	79.52%	36.03%	31.14%
M2	96.81%	91.15%	60.04%	26.21%	24.83%
M3	91.56%	90.74%	85.52%	76.57%	61.03%
M4	92.18%	91.14%	83.71%	66.20%	58.35%
BNN	94.60%	82.21%	63.45%	41.79%	35.54%
RS	93.94%	85.15%	66.98%	50.07%	35.24%

Table 3: Classification accuracy on Fashion MNIST under different levels of Gaussian noise (mean=0.0).

Methods	Original	Std = 0.1	Std = 0.2	Std = 0.3
M1	91.32%	84.55%	70.88%	49.80%
M2	91.94%	73.36%	65.24%	40.09%
M3	88.92%	87.53%	81.80%	70.92%
M4	83.51%	82.97%	81.73%	72.41%

Results and Analysis. While label smoothing, BNN, and RS perform competitively under mild perturbations, their accuracy drops sharply under stronger noise or combined corruptions, often falling below 50%. In contrast, QPLR models (M3 and M4) maintain substantially higher accuracy, particularly under severe noise and rotations. This suggests that soft labels derived from model uncertainty (BNN) or input perturbations (RS) do not provide the same robustness as QPLR's quantum-derived labels.

Tables 1, 2, and 3 further show that M2 offers only modest improvements over M1 and degrades rapidly as corruption severity increases, highlighting the limitations of heuristic smoothing methods that apply uniform label adjustments. In contrast, QPLR-based models demonstrate stronger robustness. M3 maintains stable performance by leveraging probabilistic supervision across all training samples, capturing richer class relationships and uncertainty. M4 further improves robustness by restricting training to high-confidence samples (91.5% of the dataset), filtering ambiguous data and focusing learning on reliable examples.

Notably, these trends hold across both MNIST and the more challenging Fashion-MNIST dataset, which contains greater intra-class variation and less distinctive visual cues. This further validates the effectiveness of our quantum-assisted probabilistic labeling framework in real-world settings where input uncertainty is common. Overall, while M1 and M2 may suffice in clean or slightly perturbed environments, M3 and M4 provide more resilient and uncertainty-aware learning mechanisms, making them better suited for applications such as autonomous driving or medical diagnosis where robustness to data variation is critical.

5.3 Quantitative Analysis

Confusion matrix. We report confusion matrices under challenging conditions where test samples are perturbed with Gaussian noise (mean = 0, standard deviation = 0.25) and rotated by 30 degrees: (a) M1, (b) M2, and (c) M3, shown in Figure 4. These matrices

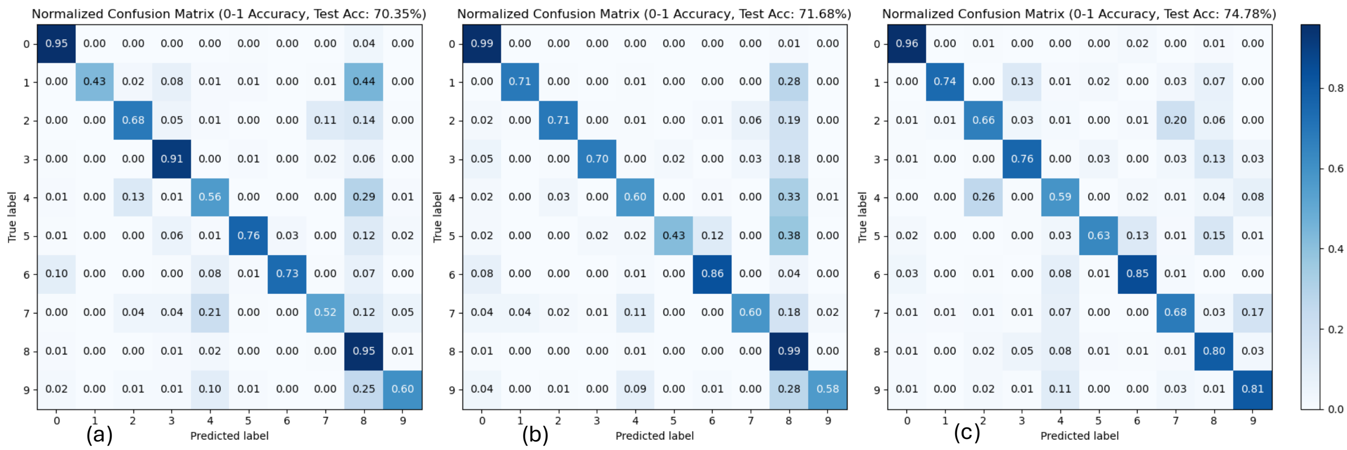


Figure 4: Confusion matrices where test samples are perturbed with Gaussian noise (mean = 0, standard deviation = 0.25) and rotated by 30 degrees: (a) M1, (b) M2, (c) M3.

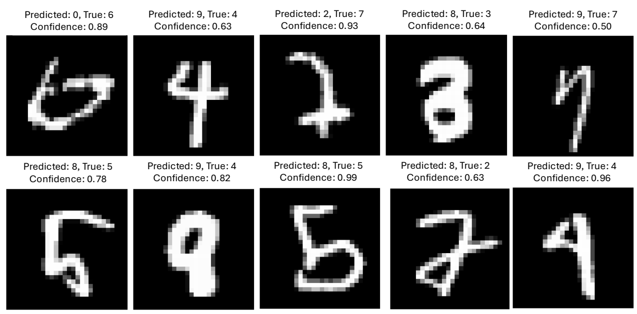


Figure 5: Samples of Wrong predictions from our M3 with all original test samples, while M1 and M2 generate predictions the same as the given labels with near-1 confidence.

provide a comprehensive view of model performance by illustrating how predictions are distributed across all classes, allowing us to assess both overall accuracy and specific misclassification trends. Compared to M1 and M2, our M3 model exhibits more balanced and accurate predictions across classes, reflecting enhanced robustness to noise and geometric distortions. This improvement underscores M3's capacity to effectively manage ambiguous or degraded inputs, an essential capability for real-world applications such as autonomous driving and medical imaging, where uncertainty and variability are common.

Label reliability. We present several challenging test samples for which our M3 model—trained with quantum-derived probabilistic labels—produces uncertain predictions that better capture the inherent ambiguity of the inputs (Figure 5). In many of these cases, even human judgment would struggle to confidently assign the correct label. In contrast, the baseline models M1 and M2 generate overconfident predictions with probabilities close to 1, strictly adhering to the given labels and overlooking the uncertainty. This distinction is particularly important in real-world applications such as autonomous driving and medical diagnosis, where recognizing and quantifying uncertainty is essential for preventing critical errors, supporting safer decisions, and fostering trust in AI-assisted systems under ambiguous or noisy conditions.

5.4 Computational Cost Comparison

To quantify the practical scalability of our hybrid quantum–classical framework, we measured wall-clock training time and final accuracy across multiple QPLR configurations (Fig. 4). All experiments were limited to 72 hours of GPU time on a single NVIDIA A100 and used the same classical pre/post network (two 128-unit dense layers) surrounding the VQC. We varied the number of qubits n (with $2^n \geq K = 10$), circuit depth L (variational layers), entanglement pattern (linear, ring, or full CZ), training epochs E, and batch size B. Increasing the qubit count from 10 to 15 increased simulation time by roughly 8× due to exponential state-vector scaling, but provided little accuracy gain on MNIST unless combined with deeper circuits and longer training. For example, a 15-qubit VQC with $L = 3$ layers trained for 15 epochs yielded a modest $\sim 2\%$ improvement, likely due to richer entanglement and extended optimization.

We observe similar scaling trends for other factors. On Fashion-MNIST, increasing circuit depth from $L = 1$ to $L = 3$ at $n = 15$ raised runtime by about 2.1× while improving accuracy by $\sim 2\%$, consistent with near-linear cost growth in circuit depth. Entanglement topology had little impact for shallow circuits, but deeper circuits with fully connected CZ gates are expected to incur higher overhead due to the increased number of two-qubit operations. Dataset complexity also affects runtime: Fashion-MNIST required roughly 25 epochs to converge, and several large configurations exceeded the 72-hour limit. These results highlight practical trade-offs when selecting VQC parameters. Emerging batched GPU quantum simulators may further reduce training cost and enable deeper circuits and longer training schedules on more challenging datasets [22, 35].

6 Conclusion

We introduced Quantum Probabilistic Label Refining (QPLR), a hybrid framework that leverages quantum non-determinism to generate soft, data-dependent labels. By exploiting the entanglement, superposition, and measurement capabilities of variational quantum circuits (VQCs), QPLR captures input uncertainty and class correlation, offering more expressive and informative labels for classical deep neural networks learning.

Table 4: Training time & performance for selected QPLR configurations. "Time" reports wall-clock hours and completed epochs.

Dataset	Encod.	Qubits	Layers	Entang.	E	B	LR	Time	Loss	Acc. (%)
MNIST	angle	10	1	linear	5	64	0.001	6h(5/5)	0.1162	96.34
MNIST	angle	10	1	linear	5	128	0.001	6h(5/5)	0.1197	96.37
MNIST	angle	10	1	linear	10	128	0.001	13h(10/10)	0.0742	97.14
MNIST	angle	15	1	linear	5	64	0.001	23h(8/15)	0.1696	95.42
MNIST	angle	15	3	linear	15	64	0.001	71h(11/15)	0.1022	97.63
FMNIST	angle	10	1	full	25	32	0.001	32h(25/25)	0.3443	88.81
FMNIST	angle	10	1	ring	25	32	0.001	32h(25/25)	0.3443	88.81
FMNIST	angle	10	1	linear	25	32	0.001	32h(25/25)	0.3443	88.81
FMNIST	angle	20	1	linear	25	32	0.001	71h(9/25)	0.3478	88.61
FMNIST	angle	20	3	linear	25	32	0.001	71h(5/25)	0.4391	84.67

Unlike traditional quantum–classical models that embed quantum layers into the training loop, QPLR decouples the quantum component by using it for one-time label generation. This offline quantum labeling allows scalable, architecture-agnostic improvements for classical models. Experiments on MNIST and Fashion-MNIST confirm that QPLR improves robustness under noise and rotation without requiring adversarial methods or architectural changes.

References

[1] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. 2021. The power of quantum neural networks. *Nature Computational Science* 1, 6 (2021), 403–409.

[2] Hamza Kamel Ahmed, Baraa Tantawi, and Gehad Ismail Sayed. 2023. Multiclass Image Classification Based on Quantum-Inspired Convolutional Neural Network.. In *CMIS*. 177–187.

[3] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. 2019. Parameterized quantum circuits as machine learning models. *Quantum science and technology* 4, 4 (2019), 043001.

[4] Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B AkashNarayanan, Ali Asadi, et al. 2018. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968* (2018).

[5] Andrew Blance and Michael Spannowsky. 2021. Quantum machine learning for particle physics using a variational quantum classifier. *Journal of High Energy Physics* 2021, 2 (2021), 1–20.

[6] Denis Bokhan, Alena S Mastiukova, Aleksey S Boev, Dmitrii N Trubnikov, and Aleksey K Fedorov. 2022. Multiclass classification using quantum convolutional neural networks with hybrid quantum-classical learning. *Frontiers in Physics* 10 (2022), 1069985.

[7] Matthias C Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J Coles. 2022. Generalization in quantum machine learning from few training data. *Nature communications* 13, 1 (2022), 4919.

[8] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. 2021. Variational quantum algorithms. *Nature Reviews Physics* 3, 9 (2021), 625–644.

[9] Avinash Chalumuri, Raghavendra Kune, and BS Manoj. 2021. A hybrid classical-quantum approach for multi-class classification. *Quantum Information Processing* 20, 3 (2021), 119.

[10] Iris Cong, Soonwon Choi, and Mikhail D Lukin. 2019. Quantum convolutional neural networks. *Nature Physics* 15, 12 (2019), 1273–1278.

[11] Chao Ding, Shi Wang, Yaonan Wang, and Weibo Gao. 2024. Quantum machine learning for multiclass classification beyond kernel methods. *arXiv preprint arXiv:2411.02913* (2024).

[12] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2021. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems* 34 (2021), 23885–23899.

[13] Biraja Ghoshal, Allan Tucker, Bal Sanghera, and Wai Lup Wong. 2021. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence* 37, 2 (2021), 701–734.

[14] Brian C Hall. 2013. *Quantum theory for mathematicians*. Springer.

[15] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. 2019. Supervised learning with quantum-enhanced feature spaces. *Nature* 567, 7747 (2019), 209–212.

[16] Max Henderson, Jarred Gallina, and Michael Brett. 2021. Methods for accelerating geospatial data processing using quantum computers. *Quantum Machine Intelligence* 3, 1 (2021), 4.

[17] Maxwell Henderson, Samriddhi Shakya, Shashindra Pradhan, and Tristan Cook. 2020. Quanvolutional neural networks: powering image recognition with quantum circuits. *Quantum Machine Intelligence* 2, 1 (2020), 2.

[18] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).

[19] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781* (2019).

[20] Katherine Hermann, Ting Chen, and Simon Kornblith. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 19000–19015.

[21] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R McClean. 2021. Power of data in quantum machine learning. *Nature communications* 12, 1 (2021), 2631.

[22] Shui Jiang, Yi-Hua Chung, Chih-Chun Chang, Tsung-Yi Ho, and Tsung-Wei Huang. 2025. BQSim: GPU-accelerated Batch Quantum Circuit Simulation using Decision Diagram. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. 79–94.

[23] Danyal Maheshwari, Daniel Sierra-Sosa, and Begonya Garcia-Zapirain. 2021. Variational quantum classifier for binary classification: Real vs synthetic dataset. *IEEE access* 10 (2021), 3705–3715.

[24] Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C Benjamin, and Xiao Yuan. 2020. Quantum computational chemistry. *Reviews of Modern Physics* 92, 1 (2020), 015003.

[25] Sam McArdle, Tyson Jones, Suguru Endo, Ying Li, Simon C Benjamin, and Xiao Yuan. 2019. Variational ansatz-based quantum simulation of imaginary time evolution. *npj Quantum Information* 5, 1 (2019), 75.

[26] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems* 34 (2021), 15682–15694.

[27] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems* 32 (2019).

[28] Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. 2022. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201* (2022).

[29] Michael A Nielsen and Isaac L Chuang. 2010. *Quantum computation and quantum information*. Cambridge university press.

[30] Wenhui Ren, Weikang Li, Shibo Xu, Ke Wang, Wenjie Jiang, Feitong Jin, Xuhao Zhu, Jiachen Chen, Zixuan Song, Pengfei Zhang, et al. 2022. Experimental quantum adversarial learning with programmable superconducting qubits. *Nature Computational Science* 2, 11 (2022), 711–717.

[31] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. 2020. Circuit-centric quantum classifiers. *Physical Review A* 101, 3 (2020), 032308.

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[33] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[34] Hao-kai Zhang, Chenghong Zhu, Mingrui Jing, and Xin Wang. 2023. Statistical analysis of quantum state learning process in quantum neural networks. *Advances in Neural Information Processing Systems* 36 (2023), 33133–33160.

[35] Yilun Zhao, Yanan Guo, Yuan Yao, Amanda Dumi, Devin M Mulvey, Shiv Upadhyay, Youtao Zhang, Kenneth D Jordan, Jun Yang, and Xulong Tang. 2022. Q-gpu: A recipe of optimizations for quantum circuit simulation using gpus. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 726–740.