SLT: ROBUST QUANTUM NEURAL NETWORKS FOR NOISY-LABEL MEDICAL IMAGE CLASSIFICATION VIA SUPERMARTINGALE-BASED LABEL TRANSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Noisy-label learning in small-scale medical image classification remains challenging, as limited data can amplify annotation errors. While quantum neural networks (QNNs) have recently been explored as compact models for small-data regimes, their behavior under label corruption remains insufficiently understood. A key obstacle is QNNs' intrinsic "natural smoothness", which may regularize training but also obscure high-confidence samples needed for noise-transition estimation. We propose Supermartingale-based Label Transition (SLT), an anchor-free loss correction framework for robust QNN-based medical image classification under noisy labels. SLT models entropy reduction in predictive distributions as a supermartingale and uses its monotonic behavior to identify stable transition-matrix refinement steps. This enables dynamic transition updates while reducing noise-driven oscillations during QNN training. We further provide a convergence analysis showing that the proposed transition-refinement process reaches a steady state. Experiments on multiple public small-scale medical image datasets demonstrate that SLT improves QNN-based noisy-label classification and outperforms classic noise-label learning baselines under synthetic and real-world label noise.

1 INTRODUCTION

Noisy-label learning in small-scale medical image classification remains a central challenge for reliable medical applications due to costly expert annotations (Shi et al., 2024). Classical deep neural networks (DNNs) have achieved remarkable success when trained on large-scale datasets with high-quality annotations (Hestness et al., 2017); however, their advantages often diminish when only limited training samples with corrupted labels are available (Song et al., 2022). This motivates alternative modeling paradigms suited to limited-data regimes. Recent studies suggest that quantum neural networks (QNNs) may offer potential benefits in small-data scenarios or tasks involving complex, high-dimensional feature spaces (Havlíček et al., 2019; Rahman & Zhuang, 2025; Oviesi et al., 2025). Nevertheless, applying QNNs to noisy-label medical image classification remains under-explored. A key challenge arises from the intrinsic "natural smoothness" of QNNs, which is rooted in the Born rule (Born, 1926). Unlike classical DNNs, whose predictive distributions can rapidly become sharply peaked, QNNs tend to produce smoother probability distributions through quantum measurements. This property can serve as an implicit regularizer in limited-data settings, but it also makes noisy-label learning more difficult: overly smooth predictions may obscure reliable class-conditional patterns and hinder the identification of high-confidence samples for noise-transition estimation. Thus, a central question remains unresolved: *can the smooth predictive behavior of* QNN*s be leveraged, rather than merely treated as a limitation, to improve robustness under noisy labels*?

To address noisy labeling issues, a line of research in classical machine learning has explored loss correction methods, which correct the training objective using a transition matrix that models the mapping between latent clean labels and observed noisy labels (Patrini et al., 2017). The effectiveness of these methods, however, heavily depends on accurate transition matrix estimation (Northcutt et al., 2021; Kye et al., 2022; Zhu et al., 2022; Liu et al., 2023). A widely adopted strategy estimates the transition matrix using anchor points, i.e., instances whose posterior distribution is nearly one-hot for a specific class, such that the class-conditional noise distribution can be directly inferred (Liu & Tao, 2015; Patrini et al., 2017). Unfortunately, true anchor points are often unreliable or even

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

unavailable in small-scale and severe label corruption medical datasets (Karimi et al., 2020; Ju et al., 2022). To mitigate this limitation, anchor-free methods have attempted to estimate the transition matrix without relying on explicit anchor points (Yao et al., 2020; Li et al., 2021; Zhu et al., 2021; Cheng et al., 2022). One representative strategy is to select high-confidence instances, also known as weak anchor points, whose predictive probabilities have low entropy and can serve as proxies for transition estimation (Xia et al., 2019; Zhang et al., 2021). While effective in some settings, such entropy-based selection can be fragile for classical DNNs due to overconfidence (Wei et al., 2022), where mislabeled samples may be selected as weak anchors, leading to biased transition estimates. In contrast, the "natural smoothness" of QNNs may reduce premature overconfidence, suggesting a promising opportunity for more stable weak-anchor selection. However, this opportunity requires a principled mechanism to determine when the transition matrix should be updated during training.

Inspired by this observation, we propose to model the entropy-reduction process of predictive distributions as a supermartingale. Rather than updating the transition matrix whenever the model becomes locally confident, our approach uses the monotonic behavior of a supermartingale to identify stable refinement steps at which predictions exhibit progressively lower entropy. This allows the transition matrix to be updated dynamically while reducing noise-driven oscillations during QNN training. Furthermore, we theoretically prove that the proposed supermartingale-based transition process converges to a steady state, providing a stability criterion for label-transition refinement. Building on this theoretical insight, we introduce a Supermartingale-based Label Transition (SLT) framework that adaptively exploits high-confidence predictions to refine the transition matrix without relying on hard-to-find anchor points. Importantly, SLT turns the smooth predictive behavior of QNNs from a potential limitation into a stabilizing inductive bias: it delays unreliable transition updates, mitigates premature overconfidence in weak-anchor selection, and improves robustness under noisy annotations. Extensive experiments across multiple public small-scale medical image datasets demonstrate that our framework improves QNN-based classification and outperforms classic noise-label learning baselines under various synthetic and real-world noise settings, where we illustrate symmetric and asymmetric noise as an example in Fig. 1. In brief, our primary contribution can be summarized as:

(a) No Noise (b) Symmetric (c) Asymmetric



Figure 1: Example probability simplex under different label-noise settings. To construct noisy labels, we flip each clean label uniformly to any other class in the symmetric case (UN), or to a fixed alternative class in a cyclic/custom manner for asymmetric cases, including cyclic-flipping (CF) and custom-mapping (CM) noise.

- We propose SLT, a new anchor-free loss correction framework for QNNs that adaptively exploits high-confidence predictions to refine the noise transition matrix without relying on hard-to-find anchor points, making it suitable for small-scale medical image datasets with noisy annotations.
- We formalize the entropy-reduction process of predictive probabilities as a *supermartingale* and theoretically analyze the convergence of the proposed transition-refinement process, providing a stability-aware foundation for label-noise correction.
- Extensive experiments on multiple public small-scale medical image datasets demonstrate that SLT improves QNN-based classification under various label noise settings, while outperforming several representative baselines.

2 PRELIMINARIES

In this section, we first introduce the preliminary background about the basics of quantum computing and barren plateaus, and then present the necessary tools from probability theory.

Learning with Noisy Labels in QNNs. In this work, we aim to train QNNs $f_{\boldsymbol{\theta}}(\cdot)$ parameterized by $\boldsymbol{\theta}$ on a noisily annotated dataset $\{\mathbf{X}, \tilde{\mathbf{Y}}\}$, where $\mathbf{X} \in \mathbb{R}^{|X| \times d}$ denotes the input matrix of $|X|$ instances with d-dimensional features, and $\tilde{\mathbf{Y}} \in \mathbb{R}^{|X| \times 1}$ denotes the observed **noisy labels**, which may deviate from the latent **ground-truth labels** $\mathbf{Y} \in \mathbb{R}^{|X| \times 1}$ (unobserved). Generally, QNNs are built by wrapping classical neural-network layers (e.g., fully connected layers) with Variational Quantum

Circuits (VQCs). Typical VQCs consist of a finite sequence of unitary blocks $U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\boldsymbol{\theta}_l)$, where $\boldsymbol{\theta}_l \in \mathbb{R}^{NR}$ collects the parameters (e.g., R rotation angles per qubit across N qubits) at layer l, and the full parameter matrix is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L) \in \mathbb{R}^{LNR}$. Each block can be written as $U_l(\boldsymbol{\theta}_l) = \exp(-i\,\boldsymbol{\theta}_l^\top \boldsymbol{V}_l)$ with a vector of Hermitian generators $\boldsymbol{V}_l$; or, in the single-parameter case, $U_l(\theta_l) = e^{-i\theta_l V_l}$ with Hermitian V_l.

To optimize the circuit, a standard variational objective $E(\boldsymbol{\theta})$ is the expectation of a Hermitian operator H under the variational state $U(\boldsymbol{\theta})|0\rangle$:

$$E(\boldsymbol{\theta}) = \langle 0|\, U(\boldsymbol{\theta})^\dagger\, H\, U(\boldsymbol{\theta})\, |0\rangle. \tag{1}$$

For K-class closed-set classification, we empirically compare the predictive probabilities $\boldsymbol{p}(\hat{Y} \mid X) \in \mathbb{R}^{|X|\times K}$, which is the measurement statistics of $U(\boldsymbol{\theta})$, with the one-hot encoded noisy labels $\tilde{\mathbf{Y}} \in \{0,1\}^{|X|\times K}$ using cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{\theta}) = -\frac{1}{|X|} \sum_{i=1}^{|X|} \sum_{k=1}^{K} \tilde{y}_{i,k}\, \log p_{i,k}, \tag{2}$$

where $p_{i,k} = p(\hat{Y} = k \mid X = x_i; \boldsymbol{\theta})$ denotes the predictive probability for the i-th instance x_i and class k, generated by the model parameterized by $\boldsymbol{\theta}$.

This formulation compares the predictive distribution with the observed noisy one-hot labels and is the standard objective for supervised classification with noisy annotations.

Loss Correction with Class-conditional Noise. Training QNNs with noisy labels inevitably risks fitting the corrupted labels rather than the latent true labels. To mitigate this issue, Patrini et al. (2017) introduced a *forward loss correction* approach that corrects the predictive probabilities $\boldsymbol{p}(\hat{Y} \mid X)$ through $\boldsymbol{p}(\tilde{Y} \mid X) = \boldsymbol{p}(\hat{Y} \mid X)\mathbf{T}$, where $\mathbf{T} \in [0,1]^{K\times K}$ is row-stochastic with entries $\mathbf{T}_{jk} = p(\tilde{Y} = k \mid \hat{Y} = j)$ for all $j, k \in \{1, \ldots, K\}$. Thus, each row of $\mathbf{T}$ describes the distribution of observed noisy labels conditioned on a predicted latent class. Due to the unobservability of latent true labels, such an approach estimates $\mathbf{T}$ using predictive probabilities $\boldsymbol{p}(\hat{Y} \mid X) \in \Delta^{K-1}$ as a proxy for true probabilities $\boldsymbol{p}(Y \mid X)$, where probability simplex $\Delta^{K-1} = \{\boldsymbol{p} \in \mathbb{R}^K \mid p_i \geq 0,\ \sum_{i=1}^{K} p_i = 1\}$. Therefore, the effectiveness of loss correction depends critically on the accuracy of $\mathbf{T}$. A common strategy constructs $\mathbf{T}$ using *anchors*, defined as:

Definition 1 (Anchor point)**.** *An instance $x^{(j)}$ is an* anchor point *for class j if $p(\hat{Y} = j \mid X = x^{(j)}) = 1$. When an anchor point $x^{(j)}$ exists for every class j, the j-th row of the noise transition matrix satisfies $\mathbf{T}_{j\cdot} = p(\tilde{Y} = \cdot \mid \hat{Y} = j)$, where $\mathbf{T}_{j\cdot} \in \Delta^{K-1}$ is a probability vector with entries $\mathbf{T}_{jk} = p(\tilde{Y} = k \mid \hat{Y} = j)$.*

The anchor concept is a common practice in prior loss-correction methods used for estimating $\mathbf{T}$; our method does not require anchors.

Tools from Probability Theory. Below, we review the definition of martingale (supermartingale), along with key tools relevant to our work.

Definition 2 (Martingale, (Williams, 1991))**.** *Let $\{M^{(t)}\}_{t\geq 1}$ be a stochastic process w.r.t. a filtration $\{\mathcal{F}^{(t)}\}_{t\geq 1}$ on a probability space $(\Omega, \mathcal{F}, P)$. The process $\{M^{(t)}\}$ is called a* martingale *if (i) $\{M^{(t)}\}$ is adapted, (ii) $\mathbb{E}[|M^{(t)}|] < \infty$, for $\forall\, t \in \mathbb{Z}^+$, (iii) $\mathbb{E}[M^{(t+1)} \mid \mathcal{F}^{(t)}] = M^{(t)}$, almost surely for $\forall\, t \in \mathbb{Z}^+$.*

If (iii) is replaced by $\mathbb{E}[M^{(t+1)} \mid \mathcal{F}^{(t)}] \leq M^{(t)}$ almost surely, we say $\{M^{(t)}\}$ is a supermartingale.

Theorem 1 (Doob's Forward Convergence Theorem, (Williams, 1991))**.** *Let $\{M^{(t)}\}_{t\geq 1}$ be an L^1-bounded* supermartingale *(in Def. 2). Then, almost surely, the limit $M^{(\infty)} = \lim_{t\to\infty} M^{(t)}$ exists and is finite.*

3 Our Proposed Framework

Our key idea is to leverage high-confidence predictions to adaptively refine the transition matrix with guaranteed convergence, without relying on hard-to-find anchor points. Prior works, such as

T-revision (Xia et al., 2019), empirically validate that estimating the transition matrix using instances with high predictive confidence can effectively improve the performance of loss correction. This **intuition** is natural: as predictive entropy decreases, the certainty of predictions increases, making them more reliable for guiding correction in training. Building on this intuition, we propose a new supermartingale-based label transition (SLT) framework that models the entropy reduction process of predictive probabilities as a supermartingale. To formulate the supermartingale process, we monitor the normalized predictive entropy (s), across $|X|$ instances and K classes:

$$s[\boldsymbol{p}(\hat{Y}|X)] = \frac{-1}{|X| \log K} \sum_{i=1}^{|X|} \sum_{k=1}^{K} p_{i,k} \log p_{i,k}. \tag{3}$$

This signal provides a global measure of model certainty: lower entropy corresponds to more confident predictions. Whenever the current s falls below its historical minimum S, we update the transition matrix accordingly. By modeling the trajectory of historical minimum S as a supermartingale, SLT leverages the monotonic accumulation of increasingly confident predictions to refine the transition matrix in an adaptive manner. The **key advantage** of this perspective lies in the favorable properties of supermartingales: *monotonicity and guaranteed convergence*. We further provide a theoretical analysis showing that the noisy transitions induced by our supermartingale formulation converge to a steady state, thereby ensuring stable loss correction.

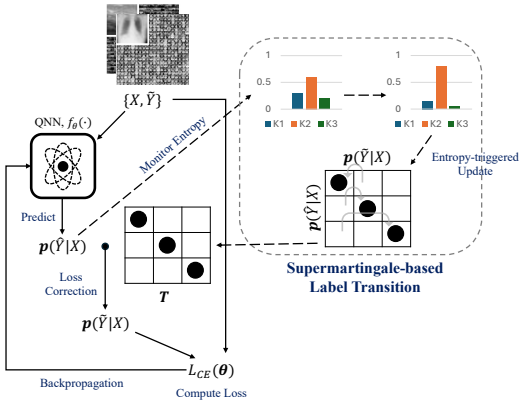

Figure 2: Overall workflow of our proposed framework, SLT. Solid arrows denote the main training process, and dashed arrows indicate the entropy-driven transition-matrix refinement.

Algorithm 1 Core steps of SLT algorithm.

Require: Noisy dataset $\{\mathbf{X}, \tilde{\mathbf{Y}}\}$, QNN $f_{\boldsymbol{\theta}}$, the number of training epoch T.

1: Initialize transition matrix $\mathbf{T}$ and score $S^{(0)}$;
2: **for** $t = 1$ to T **do**
3: $\quad \boldsymbol{p}(\hat{Y}|X) \leftarrow \mathrm{Softmax}(f_{\boldsymbol{\theta}}(\mathbf{X}))$;
4: $\quad \boldsymbol{p}(\tilde{Y}|X) \leftarrow \boldsymbol{p}(\hat{Y}|X)\mathbf{T}^{(t)}$;
5: $\quad$ Update $\boldsymbol{\theta}$ by minimizing $\mathcal{L}_{\mathrm{CE}}$;
6: $\quad$ Compute entropy score $s^{(t)}$ by Eq. (3);
7: $\quad$ **if** $s^{(t)} < S^{(t-1)}$ **then**
8: $\quad\quad$ Update $\mathbf{T}^{(t)}$ by Eqs. (4)–(5);
9: $\quad\quad$ $S^{(t)} \leftarrow s^{(t)}$;
10: $\quad$ **end if**
11: **end for**
12: **return** $\mathbf{T}$, $f_{\boldsymbol{\theta}}$.

We present the workflow of our method in Fig. 2 and provide the algorithmic details in Algo. 1. Given a noisy labeled dataset $\{\mathbf{X}, \tilde{\mathbf{Y}}\}$ and a QNN $f_{\boldsymbol{\theta}}(\cdot)$, we first initialize a warm-up transition matrix $\mathbf{T}$ (we provide implementation details in **Appendix B**) (**line 1**). After initialization, we perform T training iterations for loss correction (**line 2**). In each iteration t, we multiply (dot product) the predictive probabilities $\boldsymbol{p}(\hat{Y} \mid X) \in \mathbb{R}^{|X| \times K}$, which is generated by the QNN (**line 3**), with a transition matrix $\mathbf{T}^{(t)} \in \mathbb{R}^{K \times K}$ that models the noisy transition patterns between noisy labels and latent true labels (**line 4**). After obtaining the corrected predictive probabilities $\boldsymbol{p}(\tilde{Y} \mid X)$, we compute the cross-entropy loss between the corrected predictions and noisy labels by Eq.(2), and then update the model parameters via a gradient step with learning rate α (**line 5**). To monitor the entropy reduction process, we calculate the normalized predictive entropy $s^{(t)}$ by Eq. (3) (**line 6**). If $s^{(t)}$ is smaller than its historical minimum $S^{(t-1)}$ from previous iterations (**line 7**), indicating that the predictions have become more confident, we update the transition matrix $\mathbf{T}^{(t)}$ (**line 8**) and the historical minimum entropy $S^{(t)}$ (**line 9**). Finally, we output the refined transition matrix $\mathbf{T}$ together with the corrected QNN (**line 12**). We further analyze time and space complexity in **Appendix A**.

Dynamic refinement of the transition matrix. To refine the transition matrix $\mathbf{T}$ adaptively, at each step, we construct the dynamic transition matrix $\mathbf{T}'$ based on the current predicted labels (with lowest entropy) and the noisy labels $\tilde{Y}$. For each class pair (j, k), the entry $\mathbf{T}'[j, k]$ is defined as the

normalized co-occurrence between predicted labels $\hat{y} = j$ and noisy labels $\tilde{y} = k$:

$$\mathbf{T}'[j,k] = \frac{\sum_{i=1}^{|X|} \mathbf{1}\{\hat{y}_i = j,\, \tilde{y}_i = k\}}{\sum_{i=1}^{|X|} \mathbf{1}\{\hat{y}_i = j\}}, \tag{4}$$

where $\mathbf{1}(\cdot)$ denotes an indicator function. Intuitively, $\mathbf{T}'[j,k]$ estimates the probability that an instance predicted as class j is labeled as noisy class k.

To mitigate fluctuations due to frequent refinement, we update the global transition matrix $\mathbf{T}$ in a moving average manner as:

$$\mathbf{T} \leftarrow (1-\eta)\mathbf{T} + \eta\mathbf{T}', \tag{5}$$

where $\eta \in [0,1]$ is a step size controlling the trade-off between stability and adaptability. A smaller η emphasizes stability by retaining more of the historical matrix, while a larger η accelerates adaptation to the latest predictions.

These update rules enable SLT to progressively refine the transition matrix using increasingly confident predictions, while preserving robustness against stochastic noise.

Theoretical analysis of our framework. We first present the necessary results and further discuss their implications.

By Eq. (3), we introduce a random process based on the historical minimum of s. Let $s^{(t)}$ be the normalized predictive entropy computed at the t-th iteration. We define the historical minimum value up to iteration t as $S^{(t)} := \min_{1\le\tau\le t} s^{(\tau)} = \min\{s^{(0)}, s^{(1)}, ..., s^{(t)}\}$.

From this definition, $S^{(t)}$ is a monotonically non-increasing sequence. We can express the change at the t-th step as $S^{(t)} = S^{(t-1)} - \Delta^{(t)}$, where $\Delta^{(t)} := S^{(t-1)} - S^{(t)} = \left(S^{(t-1)} - s^{(t)}\right)_+ \ge 0$ and $(\cdot)_+$ denotes $\max(\cdot, 0)$. Unrolling this recursion yields the explicit form $S^{(t)} = S^{(0)} - \sum_{\tau=1}^{t} \Delta^{(\tau)}$.

Next, we will show that the process $\{S^{(t)}\}_{t\ge1}$ forms a supermartingale, which is a key step towards proving the convergence of our method.

Lemma 1 (Supermartingale Property)**.** *Let $\{S^{(t)}\}_{t\ge1}$ be a sequence of random variables and the natural filtration $\mathcal{F}^{(t)} = \sigma\left(S^{(0)}, \Delta^{(1)}, \cdots, \Delta^{(t)}\right)$. Then, the sequence $\{S^{(t)}\}_{t\ge1}$ is a supermartingale with respect to the filtration $\{\mathcal{F}^{(t)}\}_{t\ge1}$.*

Theorem 2 (Convergence of the Transition)**.** *The sequence of transition matrices $\{\mathbf{T}^{(t)}\}_{t\ge1}$ in Algo. 1 converges almost surely to a fixed-point matrix $\mathbf{T}^*$.*

Thm. 2 establishes convergence of the update dynamics, thereby theoretically ensuring the stability of our proposed method. We provide full proof in **Appendix A**. Notably, our framework aims to improve classification and training stability of QNNs under noisy labels, rather than forcing predictions to approach the unobserved clean labels. Since overly aggressive fitting to latent clean labels can increase overfitting risk, this work prioritizes stability-aware training under label noise.

4 EXPERIMENT

In this section, we first introduce the experimental settings and further present our results in detail.

Dataset. We first validate our proposed framework on five MedMNIST datasets: **Breast**MNIST, **Pneumonia**MNIST, **Retina**MNIST, **Derma**MNIST, and **Blood**MNIST (Yang et al., 2021). We select these datasets as they reflect the bottleneck of medical images: relatively small scale due to expensive data acquisition and annotation. This setting is more challenging than large-scale natural image datasets, since limited data magnifies the adverse effect of noisy labels, making robust learning particularly difficult. Data statistics, noise settings, and additional validations under instance-dependent noise (IDN) are presented in **Appendix B**. Furthermore, we validate our method on CheXpert (Irvin et al., 2019), a real-world dataset with noisy labels.

Implementation details. We employ a backbone QNN consisting of 8 qubits, 2 layers, and 3 rotation gates. The training process involves a 100-epoch warmup phase to build the transition matrix, followed by 100 epochs of training for loss correction. For optimization, we use the Adam optimizer with a learning rate of 0.01 and a batch size of 128. All results are reported as the mean and standard

Table 1: Comparison with baselines (CE, Forward) and competing methods on Breast, Pneumonia, and Retina, under three noise settings (UN: Uniform, CF: Cyclic-flipping, CM: Custom-mapping) at noise levels 10%/30%/50%. We report mean(std) of F1 (%) over five runs.

Noise	Models	Breast			Pneumonia			Retina		
		10%	30%	50%	10%	30%	50%	10%	30%	50%
UN	CE	66.45 (3.54)	61.21 (5.07)	47.19 (2.71)	84.87 (2.56)	83.69 (4.26)	60.82 (2.85)	27.39 (0.73)	25.14 (2.47)	22.71 (3.25)
	Forward	67.07 (2.17)	64.33 (6.12)	48.00 (2.56)	85.18 (1.88)	84.12 (1.89)	81.32 (0.67)	28.77 (1.78)	26.95 (3.18)	24.51 (2.68)
	T-Revision	64.82 (3.81)	60.06 (4.95)	43.86 (3.68)	74.24 (2.93)	68.46 (5.20)	66.22 (6.34)	21.71 (3.09)	20.60 (2.17)	20.17 (3.84)
	Dual-T	67.06 (1.23)	64.30 (5.74)	49.25 (2.53)	85.64 (1.64)	85.27 (2.85)	81.71 (1.41)	28.77 (1.78)	26.95 (3.18)	24.51 (2.68)
	VolMinNet	66.71 (1.52)	63.95 (3.41)	47.43 (2.15)	85.02 (1.24)	84.60 (1.82)	76.00 (2.51)	28.54 (1.80)	25.80 (2.17)	23.54 (3.59)
	TVR	68.06 (1.56)	61.74 (3.50)	50.61 (3.29)	76.65 (2.38)	76.93 (2.32)	68.62 (3.80)	20.90 (6.73)	18.09 (5.81)	15.91 (4.18)
	BLTM	67.31 (4.36)	59.42 (5.29)	45.02 (2.76)	83.59 (2.32)	83.21 (3.87)	65.92 (8.62)	24.27 (1.98)	21.61 (4.69)	17.25 (8.35)
	CCR	67.85 (2.19)	62.26 (3.89)	51.48 (3.48)	84.86 (2.71)	85.16 (1.57)	79.54 (2.54)	27.68 (2.86)	27.03 (2.02)	24.20 (3.50)
	Ours	**69.14 (2.71)**	**66.04 (4.34)**	**52.24 (4.41)**	**86.31 (1.63)**	**86.27 (1.81)**	**83.90 (2.09)**	**33.71 (2.93)**	**28.29 (2.23)**	**25.57 (2.88)**
CF	CE	65.64 (4.29)	59.14 (2.91)	46.70 (3.66)	82.75 (0.77)	71.42 (2.31)	45.31 (3.87)	27.40 (3.41)	26.54 (2.55)	23.38 (2.63)
	Forward	66.29 (2.01)	60.82 (5.02)	49.87 (2.88)	83.13 (2.24)	79.46 (2.72)	43.23 (5.28)	29.66 (1.54)	29.46 (1.56)	21.03 (2.98)
	T-Revision	52.18 (1.10)	51.58 (2.45)	43.73 (4.43)	78.46 (1.02)	69.46 (2.53)	36.22 (3.54)	24.48 (0.75)	22.46 (1.25)	21.46 (1.64)
	Dual-T	66.17 (3.55)	60.51 (3.16)	48.17 (4.61)	83.90 (2.43)	79.34 (1.89)	35.78 (6.82)	25.45 (1.66)	24.93 (5.33)	18.13 (7.53)
	VolMinNet	65.89 (2.31)	60.68 (2.59)	46.27 (5.31)	84.37 (1.47)	79.50 (2.74)	34.44 (7.50)	25.62 (1.96)	23.81 (5.69)	17.49 (7.77)
	TVR	64.18 (2.18)	59.76 (6.17)	43.21 (4.97)	75.28 (2.65)	55.29 (2.67)	44.30 (5.74)	18.47 (4.46)	19.96 (6.60)	19.03 (7.43)
	BLTM	67.91 (3.79)	58.45 (5.29)	49.63 (2.51)	81.26 (1.59)	73.75 (4.23)	46.62 (2.59)	25.41 (4.15)	25.31 (5.33)	17.18 (7.04)
	CCR	66.85 (1.36)	60.63 (3.42)	47.26 (4.46)	82.97 (2.86)	80.16 (3.12)	37.93 (6.89)	28.12 (2.11)	27.27 (3.64)	20.92 (3.55)
	Ours	**71.22 (2.89)**	**63.53 (3.82)**	**51.18 (3.29)**	**85.53 (1.30)**	**81.21 (2.79)**	**47.44 (5.78)**	**31.84 (3.66)**	**30.96 (2.67)**	**28.40 (3.63)**
CM	CE	65.64 (4.29)	59.14 (2.91)	46.70 (3.66)	82.75 (0.77)	71.42 (2.31)	45.31 (3.87)	27.77 (1.08)	26.72 (1.89)	22.83 (1.61)
	Forward	66.29 (2.01)	60.82 (5.02)	49.87 (2.88)	83.13 (2.24)	79.46 (2.72)	43.23 (5.28)	29.44 (2.32)	27.39 (2.53)	21.17 (2.94)
	T-Revision	52.18 (1.10)	51.58 (2.45)	43.73 (4.43)	78.46 (1.02)	69.46 (2.53)	36.22 (3.54)	23.97 (4.58)	19.25 (4.73)	18.62 (5.48)
	Dual-T	66.17 (3.55)	60.51 (3.16)	48.17 (4.61)	83.90 (2.43)	79.34 (1.89)	35.78 (6.82)	26.18 (2.61)	27.39 (2.53)	21.17 (2.94)
	VolMinNet	65.89 (2.31)	60.68 (2.59)	46.27 (5.31)	84.37 (1.47)	79.50 (2.74)	34.44 (7.50)	26.29 (3.62)	25.57 (0.97)	20.33 (3.65)
	TVR	64.18 (2.18)	59.76 (6.17)	43.21 (4.97)	75.28 (2.65)	55.29 (2.67)	44.30 (5.74)	21.24 (4.60)	20.04 (7.91)	21.34 (5.63)
	BLTM	67.91 (3.79)	58.45 (5.29)	49.63 (2.51)	81.26 (1.59)	73.75 (4.23)	46.62 (2.59)	23.78 (5.57)	20.97 (5.24)	18.04 (4.85)
	CCR	66.85 (1.36)	60.63 (3.42)	47.26 (4.46)	82.97 (2.86)	80.16 (3.12)	37.93 (6.89)	27.90 (1.23)	24.58 (5.99)	19.65 (3.90)
	Ours	**71.22 (2.89)**	**63.53 (3.82)**	**51.18 (3.29)**	**85.53 (1.30)**	**81.21 (2.79)**	**47.44 (5.78)**	**30.56 (2.78)**	**29.94 (1.77)**	**26.44 (3.05)**

deviation over five independent runs using five distinct random seeds (from 42 to 46) to ensure reproducibility. The optimal hyperparameters for our framework are reported in Tab. 8. We report classification performance using the macro-averaged F1 score, as class imbalance is prevalent in medical image datasets. A higher F1 score implies a better classification.

Baseline and competing methods. For fair comparison, we compare our framework with the standard baseline and competing methods that are widely used as anchor-free label transition approaches. **CE** denotes the approach that we directly train the model with cross-entropy loss without loss correction. **Forward** loss correction (Patrini et al., 2017) attempts to correct predictive probabilities with an estimated transition matrix. **T-Revision** (Xia et al., 2019) builds the initial transition matrix from high-confident predicted instances and then iteratively refines it with a learnable slack term ΔT while jointly optimizing the classifier by minimizing the weighted loss. **Dual-T** (Yao et al., 2020) jointly leverages forward- and backward-corrected losses to more accurately estimate the noise transition matrix. **VolMinNet** (Li et al., 2021) proposes an anchor-free framework that regularizes the cross-entropy loss with the volume of the simplex formed by the columns of the transition matrix. **TVR** (Zhang et al., 2021) applies total variation regularization to identify "cleanest" predictive probabilities without explicitly detecting anchor points. **BLTM** (Yang et al., 2022) leverages Bayes-optimal labels inferred from distilled clean samples to better estimate the transition matrix. **CCR** (Cheng et al., 2022) estimates the transition matrix using a forward-backward cycle-consistency regularization without relying on anchor points. We provide hyperparameter settings in **Appendix B**.

Our framework performs robustly against baselines and competing methods under three types of label noise across diverse datasets. We evaluate our framework (Ours) against two baselines (Cross-Entropy, i.e., CE, and Forward) and six representative anchor-free label-transition-based methods (T-Revision, Dual-T, VolMinNet, TVR, BLTM, CCR) on five MedMNIST datasets (Breast, Pneumonia, Retina, Derma, Blood) under three types of noise with varying noise ratios (10%, 30%, 50%). All experiments are repeated five times, and we report the mean $\pm$ std of the F1 score. As shown in Tab. 1, our method outperforms competing approaches across most datasets and noise levels. For example, on Breast and Pneumonia, our framework maintains a clear advantage even under severe noise (50%), achieving 52.24% and 83.90%, respectively, whereas conventional baselines such as CE

and Forward suffer from substantial performance degradation. These results demonstrate that our framework achieves superior stability and robustness across diverse datasets and noise conditions, effectively mitigating the adverse effects of label noise on QNNs' performance. Due to the page limit, we present the results of Derma and Blood in **Appendix B**.

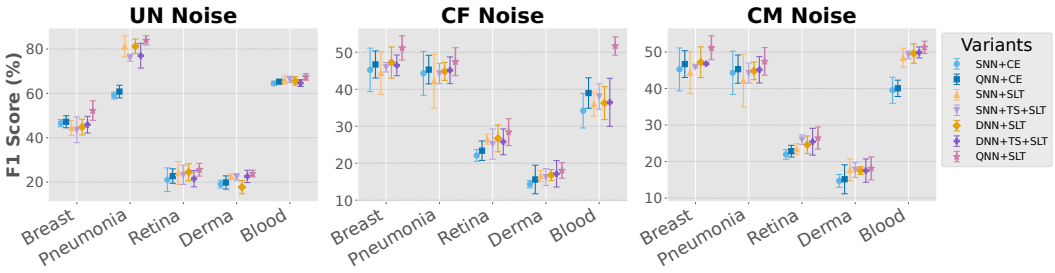

Figure 3: Ablation study on VQCs evaluated by F1 score (mean ± std) across five datasets under three types of noises, uniform (UN), cyclic-flipping (CF), and custom-mapping (CM) noise, with 50% noise ratio. We compare our backbone QNN with classical shallow and deep neural networks (NNs) on two training schemes, direct training using cross-entropy (CE) and our framework (SLT), respectively. To assess how smoothness affects NNs' output, we apply temperature scaling (TS) to both NNs before employing SLT.

Ablation studies on backbone models. We conduct additional ablation studies to evaluate the effectiveness of the QNNbackbone under three types of noise at a 50% noise ratio across five datasets. We first compare SLT with CE-only training using both classical and quantum models, and then evaluate three SLT backbones: a two-layer shallow neural network (SNN), a classical deep neural network (DNN) implemented as a ResNet (He et al., 2016) with ten residual blocks, and a QNN. As shown in Fig. 3, QNNconsistently demonstrates advantages over classical backbones on datasets with limited instances and noisy annotations. Besides, we observe that using classical models as the backbone of SLT may lead to performance degradation. We argue that classical models are prone to severe early-stage overconfidence (Wei et al., 2022), often pushing Softmax outputs to extremes and causing predictive entropy to collapse to nearly zero even under noisy labels. This problem is further amplified in our framework, since SLT relies on entropy reduction to update the transition matrix. To further examine this overconfidence issue, we apply temperature scaling (TS) (Guo et al., 2017) to both SNN and DNN. We fine-tune the temperature on the validation set and select 1.5 in this setting. The results show that although DNN is generally slightly better than SNN in most cases, applying TS does not consistently improve the performance of either backbone in SLT. In contrast, QNNs exhibit an inherent "natural smoothness" rooted in the Born rule (Born, 1926), where outputs are derived from projection probabilities of quantum states governed by unitary transformations and limited Hilbert space expressivity; this prevents the probability distribution from instantaneously collapsing to a one-hot state. This gradual growth of confidence matches SLT's monotonic yet steady entropy reduction process well, allowing our framework to capture the true learning trajectory from uncertainty to certainty rather than being misled by the deceptive over-confident signals often seen in classical models. This observation further suggests that SLT has particular advantages when combined with QNNs. Furthermore, we analyze the trainable parameters for backbone models and discuss how the smoothness affects QNNs in **Appendix B**.

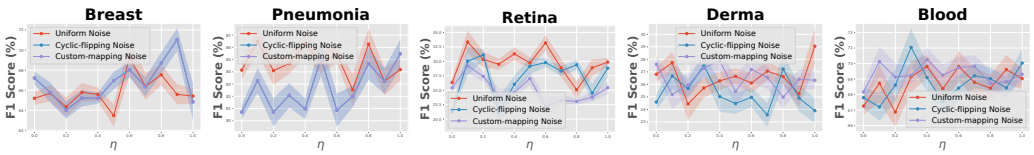

Figure 4: Sensitivity analysis of hyperparameters η under three noise types for five datasets. We run the experiments five times and present the mean F1 score with standard deviation in shaded regions.

Analysis of hyperparameters η. We analyze the sensitivity of hyperparameters η under different noise types (Uniform, Cyclic-flipping, and Custom-mapping) with a 10% noise ratio for five datasets on the validation set. The transition matrix is updated with the "NDU+PTU" mode (0.5 del, 10 pat).

The update strategies are discussed in **Appendix B**. Each curve represents the average F1 score, with the shaded area denoting the standard deviation across five runs. We observe the patterns of η in Fig. 4. Overall, the curves show that the optimal η varies across datasets. For asymmetric noises such as Cyclic-flipping and Custom-mapping, the performance is identical on Breast and Pneumonia due to binary classification tasks, where pair-wise label flips produce the same effect. Pneumonia tends to favor large values of η under pair-wise noise, while more challenging datasets (Retina and Derma) achieve better results at either very small or extreme settings. Notably, although determining the optimal η requires hyperparameter tuning, the performance degradation under sub-optimal settings is generally mild, thereby validating the robustness of our method.

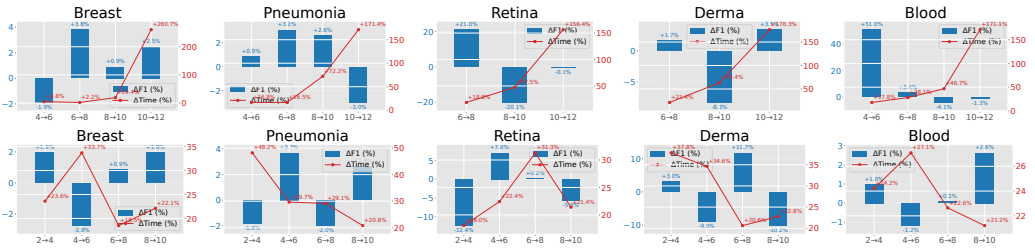

Figure 5: Trade-off analysis under different numbers of qubits (upper row) and layers (lower row). We report incremental F1 gain (%) in blue bars and Time cost (%) in red lines as the number of consecutive qubits/layers increases.

Trade-off Analysis under different numbers of qubits and layers. We first analyze the performance–cost trade-off of QNNs on the validation set and report incremental changes in F1 (blue bars) and runtime (red line) as the number of qubits increases from 4 to 12 while fixing the depth to two layers. To isolate the effect of qubit count, we perform a controlled ablation where only the number of qubits varies and all other settings are held constant across runs. The QNN is evaluated under symmetric noise at a 10% rate and trained with the "NDU+PTU" strategy (0.5 del, 10 pat). In the upper row of Fig. 5, we observe that moving from 6 to 8 qubits consistently offers the best performance–cost trade-off, whereas further increases show clear diminishing (or even negative) returns with sharp cost inflation. For example, on Breast, employing 8 qubits yields 3.81% F1 gain for only 2.18% runtime increase. In contrast, increasing the qubits from 8 to 10 just obtains 0.90% F1 gain at an additional 19.37% runtime cost. Notably, for Retina and Derma, a valid warm-up transition matrix fails to be built, i.e., predicted classes mismatched the classes in noisy labels, so that these cases are excluded in comparison. Taken together across datasets, 8 qubits emerges as the global "sweet spot": before 8, F1 improves with moderate cost; beyond 8, F1 plateaus or declines while runtime escalates steeply. We therefore choose 8 qubits for this study. Besides, we fix the number of qubits as 8 and vary the number of layers from 2 to 10. As the number of layers increases, the F1 score increases moderately, but the time cost increases hugely. Considering that we aim to enhance the training under noise-label setting with a cost-effective budget, we choose two layers so we can efficiently validate our methods.

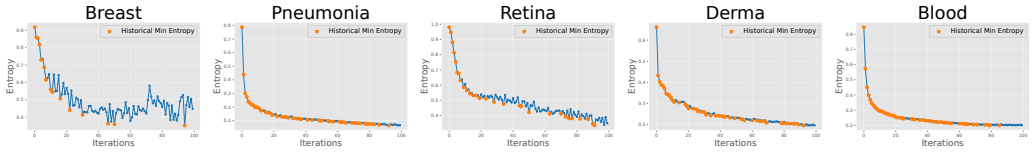

Figure 6: Entropy reduction process. We mark the running historical lowest entropy in orange.

Discussion on the entropy reduction process. We present the entropy reduction process across five datasets in Fig. 6. Historical minima up to the past iterations are highlighted in orange. The results validate that the entropy reduction process is effective across all datasets. On Breast, the entropy shows larger oscillations, likely due to the limited number of instances, which amplifies predictive uncertainty. The remaining datasets exhibit a two-stage behavior, with a rapid drop in early iterations as high-uncertainty instances are corrected, followed by a slower decay as the model refines decision boundaries around hard cases. Overall, these trends demonstrate that our method consistently reduces uncertainty in the predictive distribution.

Visualization of noisy transition. Fig. 7 visualizes the evolution of the class-dependent noise transition (10% symmetric noise) on Blood. Although early iterations contain noticeable off-diagonal mass, diagonal entries increase while off-diagonal probabilities are compressed and become sparse as noise transits, indicating progressive noise suppression. This evolution aligns with our iterative estimation of the transition matrix, which leverages high-confidence predictions to correct the noise distribution.

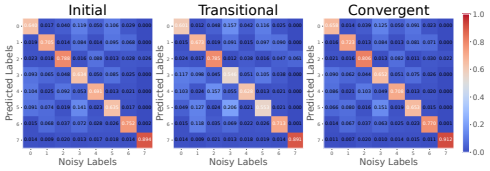

Figure 7: Example of noisy transition process (10% symmetric noise). We present the initial state, intermediate transition state, and the final convergence of the transition matrix on Blood.

5 Related Work

Noisy-label learning. Learning with noisy labels has been widely studied in classical deep learning models through robust losses (Liu & Guo, 2020; Feng et al., 2022; Liu et al., 2022), memorization-based sample selection (Han et al., 2018; Chen et al., 2019; Wang et al., 2022), label correction (Tanaka et al., 2018; Ji et al., 2021; Algan & Ulusoy, 2022), co-training or co-regularization (Wei et al., 2020; Sheng et al., 2025), and contrastive representation learning (Li et al., 2022; Karim et al., 2022; Wu et al., 2020), under class-dependent (Liu & Tao, 2015; Patrini et al., 2017; Xia et al., 2019; Yao et al., 2020; Kye et al., 2022; Zhu et al., 2022; Liu et al., 2023) and instance-dependent noise (Yao et al., 2023; Xia et al., 2023; Yang et al., 2024; Nguyen et al., 2024; Li et al., 2025). A prominent approach models class-conditional corruption with a noise transition matrix and performs forward or backward loss correction (Patrini et al., 2017). Early estimators rely on anchor points (Liu & Tao, 2015) with near one-hot class posteriors, an assumption fragile under small and highly corrupted datasets. Anchor-free methods instead iteratively refine the transition using confident predictions or structural regularization (Xia et al., 2019; Yao et al., 2020; Li et al., 2021; Zhang et al., 2021; Zhu et al., 2021; Yang et al., 2022; Cheng et al., 2022). However, many still rely on high-confidence selection and lack convergence guarantees for transition refinement. Our proposed framework adaptively selects pseudo-anchors using a supermartingale criterion on predictive entropy and stabilizes transition updates via moving averaging, yielding convergence-guaranteed refinement suited to small-data settings such as medical imaging.

Quantum neural networks for medical image classification. QNNs have shown potential in small-data regimes (Jia et al., 2019; Hur et al., 2022; Moussa et al., 2024), including medical imaging (Mathur et al., 2021; Landman et al., 2022), where high-quality annotations are often scarce (Wei et al., 2023). Existing approaches include pure quantum circuits (Yousif et al., 2024), quantum-classical hybrid models (Mathur et al., 2021; Ajlouni et al., 2023), and quantum-inspired models (Ibrahim et al., 2024; Li et al., 2026). Yet systematic treatment of label noise remains under-explored for QNNs (Ju et al., 2022). We address this gap with a supermartingale-based label transition mechanism for robust loss correction and QNN training under label noise.

6 Conclusion

We introduce Supermartingale-based Label Transition (SLT), an anchor-free loss-correction framework for robust QNN training under noisy labels in medical imaging. SLT treats the historical decrease of predictive entropy as a supermartingale process and uses this signal to trigger stable transition-matrix refinement, avoiding reliance on hard-to-find anchor points and providing a theoretical guarantee to steady convergence. Experiments on multiple small-scale medical image datasets demonstrate that SLT improves QNN-based classification under various label noise settings, consistently outperforming representative baselines.

A main **limitation** is the computational cost of QNN simulation, a well-known bottleneck in Quantum Machine Learning (QML) rather than a specific drawback of our framework. **Future work** will evaluate SLT under more complex real-world noise patterns and across broader medical imaging modalities. More **broadly**, our framework has the potential to enhance robust medical image analysis, particularly for smaller clinics or healthcare centers in rural areas with limited resources to obtain high-quality medical data.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing writing and LaTeX formatting. We did not use generative AI tools to generate synthetic datasets. All content was independently checked by the authors.

ETHICS STATEMENT

This work studies robust learning from noisy labels for medical image classification using publicly available benchmark datasets, including MedMNIST and CheXpert. We did not recruit human participants, collect new patient data, or access personally identifiable information. Our experiments use existing datasets according to their established research protocols and focus on methodological evaluation rather than clinical deployment. Performance may vary across datasets, populations, imaging conditions, and annotation practices, and deployment in clinical settings would require additional validation, fairness assessment, privacy safeguards, and compliance with applicable regulatory and ethical requirements.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce the main results throughout the paper and appendix. Section 4 describes the datasets, evaluation metrics, backbone configuration, optimization settings, and repeated-run protocol. Appendix A provides complete proofs for the theoretical results, including the assumptions and derivations underlying the convergence analysis. Appendix B provides supplementary experimental details, including dataset statistics (Table 2), noise-generation procedures, the QNN architecture (Figure 8), baseline implementations, hyperparameter settings (Table 8), transition-update strategies, additional experiments, and the computational environment. All experiments are repeated over five independent random seeds, with mean and standard deviation reported. Source code will be released upon acceptance to further support reproducibility.

REFERENCES

Naim Ajlouni, Adem Özyavaş, Mustafa Takaoğlu, Faruk Takaoğlu, and Firas Ajlouni. Medical image diagnosis based on adaptive hybrid quantum cnn. *BMC Medical Imaging*, 23(1):126, 2023.

Görkem Algan and Ilkay Ulusoy. Metalabelnet: Learning to generate soft-labels from noisy-labels. *IEEE Transactions on Image Processing*, 31:4352–4362, 2022.

Max Born. Quantenmechanik der stoßvorgänge. *Zeitschrift für physik*, 38(11):803–827, 1926.

Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International conference on machine learning*, pp. 1062–1070. PMLR, 2019.

Yuanhong Chen, Fengbei Liu, Hu Wang, Chong Wang, Yuyuan Liu, Yu Tian, and Gustavo Carneiro. Bomd: bag of multi-label descriptors for noisy chest x-ray classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21284–21295, 2023.

De Cheng, Yixiong Ning, Nannan Wang, Xinbo Gao, Heng Yang, Yuxuan Du, Bo Han, and Tongliang Liu. Class-dependent label-noise learning with cycle-consistency regularization. *Advances in Neural Information Processing Systems*, 35:11104–11116, 2022.

Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. Ssr: an efficient and robust framework for learning with unknown label noise. In *33rd British Machine Vision Conference Proceedings, BMVC 2022*, 2022.

Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2288–2298, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

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

Hao-Yuan He, Yu Liu, Ren-Biao Liu, Zheng Xie, and Ming Li. Probabilistic instance dependent label refinement for noisy label learning. *Machine Learning*, 114(5):120, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Tak Hur, Leeseok Kim, and Daniel K Park. Quantum convolutional neural network for classical data classification. *Quantum Machine Intelligence*, 4(1):3, 2022.

Hussain K Ibrahim, Nizar Rokbani, Ali Wali, Khmaies Ouahada, Habib Chabchoub, and Adel M Alimi. A medical image classification model based on quantum-inspired genetic algorithm. *Engineering, Technology & Applied Science Research*, 14(5):16692–16700, 2024.

Sontje Ihler, Felix Kuhnke, Timo Kuhlgatz, and Thomas Seel. Distribution-aware multi-label fixmatch for semi-supervised learning on chexpert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2295–2304, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 105–115, 2021.

Daehyun Ji, Dokwan Oh, Yoonsuk Hyun, Oh-Min Kwon, and Myeong-Jin Park. How to handle noisy labels for robust learning from uncertainty. *Neural networks*, 143:209–217, 2021.

Zhih-Ahn Jia, Biao Yi, Rui Zhai, Yu-Chun Wu, Guang-Can Guo, and Guo-Ping Guo. Quantum neural network states: A brief review of methods and applications. *Advanced Quantum Technologies*, 2 (7-8):1800077, 2019.

Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu, and Zongyuan Ge. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging*, 41(6):1533–1546, 2022.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9676–9686, 2022.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, 65:101759, 2020.

Seong Min Kye, Kwanghee Choi, Joonyoung Yi, and Buru Chang. Learning with noisy labels by efficient transition matrix estimation to combat label miscorrection. In *European Conference on Computer Vision*, pp. 717–738. Springer, 2022.

Jonas Landman, Natansh Mathur, Yun Yvonna Li, Martin Strahm, Skander Kazdaghli, Anupam Prakash, and Iordanis Kerenidis. Quantum methods for neural networks and application to medical image classification. *Quantum*, 6:881, 2022.

Dongfen Li, Yuchen Sun, Yuhang Yuan, Zhikang Hu, Qiuyu Xiang, Yangyang Jiang, Yonghao Zhu, You Fu, and Xiaoyu Hua. A quantum-inspired medical scalable convolutional neural network for intelligent pneumonia diagnosis. *Biomedical Signal Processing and Control*, 112:108440, 2026.

Jiahui Li, Tai-Wei Chang, Kun Kuang, Ximing Li, Long Chen, and Jun Zhou. Learning causal transition matrix for instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18305–18313, 2025.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 316–325, 2022.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *International conference on machine learning*, pp. 6403–6413. PMLR, 2021.

Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24070–24079, 2023.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pp. 14153–14172. PMLR, 2022.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International conference on machine learning*, pp. 6226–6236. PMLR, 2020.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *International Conference on Machine Learning*, pp. 21475–21496. PMLR, 2023.

Yuan Ma, Junlin Hou, Chao Zhang, Yukun Zhou, Zongyuan Ge, Haoran Xie, and Lie Ju. Benchmarking real-world medical image classification with noisy labels: Challenges, practice, and outlook. *arXiv preprint arXiv:2512.09315*, 2025.

Natansh Mathur, Jonas Landman, Yun Yvonna Li, Martin Strahm, Skander Kazdaghli, Anupam Prakash, and Iordanis Kerenidis. Medical image classification via quantum neural networks. *arXiv preprint arXiv:2109.01831*, 2021.

Charles Moussa, Yash J Patel, Vedran Dunjko, Thomas Bäck, and Jan N Van Rijn. Hyperparameter importance and optimization of quantum neural networks across small datasets. *Machine Learning*, 113(4):1941–1966, 2024.

Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Noisy label learning with instance-dependent outliers: Identifiability via crowd wisdom. *Advances in Neural Information Processing Systems*, 37:97261–97298, 2024.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

Safura Oviesi, Mohamad Jafar Tarokh, and Mohamad kazem Momeni. Quantum neural network-assisted learning for small medical datasets: a case study in emphysema detection. *The Journal of Supercomputing*, 81(1):308, 2025.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Maqsudur Rahman and Jun Zhuang. Nqnn: Noise-aware quantum neural networks for medical image classification. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15972. Springer Nature Switzerland, September 2025.

Mengmeng Sheng, Zeren Sun, Tianfei Zhou, Xiangbo Shu, Jinshan Pan, and Yazhou Yao. Ca2c: A prior-knowledge-free approach for robust label noise learning via asymmetric co-learning and co-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 901–911, 2025.

Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. A survey of label-noise deep learning for medical image analysis. *Medical image analysis*, 95:103166, 2024.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5552–5560, 2018.

Yikai Wang, Xinwei Sun, and Yanwei Fu. Scalable penalized regression for noise detection in learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 346–355, 2022.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13726–13735, 2020.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pp. 23631–23644. PMLR, 2022.

Lin Wei, Haowen Liu, Jing Xu, Lei Shi, Zheng Shan, Bo Zhao, and Yufei Gao. Quantum machine learning in medical image analysis: A survey. *Neurocomputing*, 525:42–53, 2023.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. *Advances in neural information processing systems*, 33: 21382–21393, 2020.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in neural information processing systems*, 33:7597–7610, 2020.

Xiaobo Xia, Bo Han, Yibing Zhan, Jun Yu, Mingming Gong, Chen Gong, and Tongliang Liu. Combating noisy labels with sample selection by mining high-discrepancy examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1843, 2023.

Ruixuan Xiao, Yiwen Dong, Haobo Wang, Lei Feng, Runze Wu, Gang Chen, and Junbo Zhao. Promix: combating label noise via maximizing clean sample utility. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4442–4450, 2023.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

Jie Yang, Xiaoguang Niu, Yuanzhuo Xu, Zejun Zhang, Guangyi Guo, Steve Drew, and Ruizhi Chen. Dynamic selection for reconstructing instance-dependent noisy labels. *Pattern Recognition*, 156: 110803, 2024.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *International Conference on Machine Learning*, pp. 25302–25312. PMLR, 2022.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020.

Yu Yao, Mingming Gong, Yuxuan Du, Jun Yu, Bo Han, Kun Zhang, and Tongliang Liu. Which is better for learning with noisy labels: the semi-supervised method or modeling label noise? In *International conference on machine learning*, pp. 39660–39673. PMLR, 2023.

Mohammed Yousif, Belal Al-Khateeb, and Begonya Garcia-Zapirain. A new quantum circuits of quantum convolutional neural network for x-ray images classification. *IEEE access*, 12:65660–65671, 2024.

Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *International conference on machine learning*, pp. 12501–12512. PMLR, 2021.

Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, pp. 12912–12923. PMLR, 2021.

Zhaowei Zhu, Jialu Wang, and Yang Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning*, pp. 27633–27653. PMLR, 2022.

In the appendix, we first provide the full proof for our theoretical analysis, and further exhibit supplementary experimental settings (e.g., noise design and model architecture) and results (e.g., update strategies). The source code will be released upon acceptance to support reproducibility.

A THEORETICAL ANALYSIS

Full proof of the theoretical analysis. We provide full proof of our theoretical analysis as follows.

Proof for Lemma 1. According to Def. 2, a process $S^{(t)}$ is a supermartingale relative to $(\Omega, \mathcal{F}, P)$ if it satisfies Adaptedness, Integrability, and Supermartingale.

Adaptedness. We first aim to verify that $S^{(t)}$ is determined based on the information available up to past t iterations. $S^{(t)} = S^{(0)} - \sum_{\tau=1}^{t} \Delta^{(\tau)}$ is a finite subtraction of random variables that are measurable w.r.t. $\sigma\big(\Delta^{(1)}, \ldots, \Delta^{(t)}\big)$. Thus, $S^{(t)}$ is also measurable w.r.t. $\mathcal{F}^{(t)}$, ensuring the adaptedness.

Integrability. From Eq. (3), the normalized predictive entropy s lies in $[0, 1]$. Consequently, its historical minimum also resides in $[0, 1]$. Hence, we obtain $\mathbb{E}[|S^{(t)}|] \leq 1 < \infty$, which guarantees integrability of $\mathbb{E}[|S^{(t)}|]$ for all t.

Supermartingale. Let $\mathbb{E}_{\mathcal{F}^{(t-1)}}[\cdot] := \mathbb{E}[\ \cdot\ |\ \mathcal{F}^{(t-1)}]$. Since $S^{(t-1)}$ is $\mathcal{F}^{(t-1)}$-measurable, $\mathbb{E}_{\mathcal{F}^{(t-1)}}[S^{(t-1)}] = S^{(t-1)}$. The derivation is as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{F}^{(t-1)}}\left[S^{(t)}\right] &= \mathbb{E}_{\mathcal{F}^{(t-1)}}\left[S^{(t-1)} - \Delta^{(t)}\right] \\
&= \mathbb{E}_{\mathcal{F}^{(t-1)}}\left[S^{(t-1)}\right] - \mathbb{E}_{\mathcal{F}^{(t-1)}}\left[\Delta^{(t)}\right] \\
&= S^{(t-1)} - \mathbb{E}_{\mathcal{F}^{(t-1)}}\left[\max(S^{(t-1)} - s^{(t)}, 0)\right] \\
&\leq S^{(t-1)} - \max(\mathbb{E}_{\mathcal{F}^{(t-1)}}\left[S^{(t-1)} - s^{(t)}\right], 0) \\
&\leq S^{(t-1)}.
\end{aligned}$$

Thus, the supermartingale condition holds true for $\forall\, t \geq 1$ s.t.

$$\mathbb{E}\big[S^{(t)}\big|\mathcal{F}^{(t-1)}\big] \leq S^{(t-1)}, \quad \forall\, t \geq 1.$$

□

Proof for Theorem 2. By Thm. 1, the limit $\lim_{t\to\infty} S^{(t)} = S^{(\infty)}$ exists and is finite. Given that $\{S^{(t)}\}$ is a monotonically non-increasing sequence, this indicates that the difference between two adjacent terms must almost surely approach 0,

$$\lim_{t\to\infty} \Delta^{(t)} = \lim_{t\to\infty} \big(S^{(t-1)} - S^{(t)}\big) = 0.$$

By Algo. 1, the condition for updating the transition matrix $\mathbf{T}$ is $s^{(t)} < S^{(t-1)}$, equivalent to $\Delta^{(t)} > 0$. As proven, $\lim_{t\to\infty} \Delta^{(t)} = 0$ (a.s.), which implies that the updates of $\mathbf{T}$ become asymptotically infrequent.

The above results indicate that for any $\epsilon > 0$, there exists a positive integer T such that $\forall s, t > T$,

$$|\mathbf{T}^{(s)} - \mathbf{T}^{(t)}| < \epsilon,$$

which implies that the sequence $\{\mathbf{T}^{(t)}\}$ is a Cauchy sequence in the space of stochastic transition matrices. Since this space is complete, the sequence $\{\mathbf{T}^{(t)}\}$ converges to a fixed-point matrix $\mathbf{T}^*$. □

Analysis of time and space complexity. Time complexity per iteration is dominated by the QNN's forward/backward propagation with parameter size $\mathcal{O}(LNR)$, which may generally take $\mathcal{O}(|X|\cdot LR\cdot 2^N)$ in classical quantum simulators, while the label correction step requires a dot product between predictive probabilities and the transition matrix, which takes $\mathcal{O}(|X|K^2)$ cost. Besides, computing

cross-entropy and predictive entropy takes $\mathcal{O}(|X|K)$, whereas updating the transition matrix takes $\mathcal{O}(|X|)$. Overall, the **time complexity** is $\mathcal{O}(|X|\cdot LR\cdot 2^N + |X|K^2)$. During training, we store the model parameters $\mathcal{O}(LNR)$, the predictive probabilities $\mathcal{O}(|X|K)$, and the transition matrix $\mathcal{O}(K^2)$. So, the **space complexity** is $\mathcal{O}(LNR + |X|K + K^2)$.

B Supplementary Experiments

Data Statistics. We summarize the data statistics of the MedMNIST datasets in Tab. 2.

Table 2: Statistics of the MedMNIST datasets used in our experiments. All images are converted to gray scale with 28×28 size. $|X_{tr}|$, $|X_{val}|$, $|X_{te}|$, $|X|$, and K denote the number of train/validation/test/total instances and classes, respectively.

Dataset	Modality	$\lvert X_{tr}\rvert$	$\lvert X_{val}\rvert$	$\lvert X_{te}\rvert$	$\lvert X\rvert$	K
Breast	Breast Ultrasound	546	78	156	780	2
Pneumonia	Chest X-Ray	4,708	524	624	5,856	2
Retina	Fundus Camera	1,080	120	400	1,600	5
Derma	Dermatoscope	7,007	1,003	2,005	10,015	7
Blood	Microscope	11,959	1,712	3,421	17,092	8

Table 3: Mapping rubric for the custom-mapping noise. Each class is paired with a clinically or visually confusable counterpart.

Dataset	**Original Classes**	**Noisy-paired Classes**
Breast	0 (malignant)	1 (benign)
	1 (benign)	0 (malignant)
Pneumonia	0 (normal)	1 (pneumonia)
	1 (pneumonia)	0 (normal)
Retina	0 (no DR)	1 (mild NPDR)
	1 (mild NPDR)	0 (no DR)
	2 (moderate NPDR)	3 (severe NPDR)
	3 (severe NPDR)	2 (moderate NPDR)
	4 (PDR)	3 (severe NPDR)
Derma	0 (AKIEC)	1 (BCC)
	1 (BCC)	0 (AKIEC)
	2 (BKL)	5 (NV)
	3 (DF)	2 (BKL)
	4 (MEL)	5 (NV)
	5 (NV)	4 (MEL)
	6 (VASC)	1 (BCC)
Blood	0 (basophil)	1 (eosinophil)
	1 (eosinophil)	0 (basophil)
	2 (erythroblast)	3 (immature granulocytes)
	3 (immature granulocytes)	6 (neutrophil)
	4 (lymphocyte)	5 (monocyte)
	5 (monocyte)	4 (lymphocyte)
	6 (neutrophil)	3 (immature granulocytes)
	7 (platelet)	2 (erythroblast)

Noise implementation and design. We follow the setting in **Forward** (Patrini et al., 2017) to construct uniform (UN) and pair-wise (Cyclic-flipping, UF) noise. Furthermore, we design a new pair-wise noise (Custom-mapping, CM) and justify the design, whose mapping rubric is provided in Tab. 3. For the binary-class datasets (**Breast** and **Pneumonia**), we use a standard pair-wise mapping between the two classes. For the complex multi-class datasets, the pairings are designed to simulate realistic, clinically plausible diagnostic errors.

For **Retina**, we leverage the ordinal and progressive nature of diabetic retinopathy (DR) classification. The dataset comprises five clinically defined stages of DR: no apparent retinopathy, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR). Misdiagnoses in real-world settings often occur between adjacent stages due to overlapping retinal features such as microaneurysms, hemorrhages, and neovascularization, especially under suboptimal imaging conditions. To simulate such realistic diagnostic ambiguities, we construct a class-conditional pair noise schema wherein each DR stage is paired with its immediate neighbor. Specifically, "no DR" is paired with "mild NPDR", "moderate NPDR" with "severe NPDR", and "severe NPDR" with "PDR". This setup reflects both clinical practice and morphological continuity.

For **Derma**, we construct a pairwise noise mapping based on dermatological similarity, visual confusion, and misdiagnosis patterns in dermoscopy. Each class is mapped to a single "confusable" counterpart, capturing asymmetric mislabeling commonly observed in medical practice. For instance, actinic keratoses and intraepithelial carcinoma (AKIEC) are frequently mistaken for basal cell carcinoma (BCC) due to overlapping visual features such as keratinization and pigmentation. Similarly, melanoma (MEL) and melanocytic nevi (NV) represent a classic diagnostic challenge, as their visual boundaries and color distributions often overlap, leading to bidirectional confusion. Benign keratosis-like lesions (BKL) can appear similar to nevi, while dermatofibroma (DF) may be misinterpreted as BKL due to shared dermal structures. Vascular lesions (VASC), characterized by reddish nodules, may occasionally mimic the appearance of BCC in dermoscopic images.

For **Blood**, each cell type is paired with another class that it is most likely to be confused with due to visual resemblance under light microscopy, particularly in low-resolution or poorly stained images. For instance, basophils and eosinophils (classes 0 and 1) are both granulocytes with similar nuclear granulation, making them difficult to differentiate. Erythroblasts (class 2), which are immature red blood cells, share size and chromatin features with immature granulocytes (class 3), especially in early developmental stages. Likewise, lymphocytes and monocytes (classes 4 and 5) are frequently misclassified due to overlapping sizes and nuclear shapes. Finally, platelets (class 7), being small and dense, may be mistaken for erythroblastic fragments under certain imaging conditions.

Model architecture of our backbone QNNs. In this study, we evaluate our framework using a QNN as our backbone model. Unlike a standalone quantum circuit, our model incorporates classical pre-processing and post-processing layers to enhance learnability. First, high-dimensional classical inputs are compressed via a fully connected layer and normalized using a hyperbolic tangent ($\tanh$) activation function. This step ensures that input values remain bounded, preventing the rotation angles in the subsequent quantum embedding from diverging. The processed features are then encoded into quantum states via angle encoding using R_X rotations. The variational ansatz employs *StronglyEntanglingLayers*, which consists of parameterized single-qubit rotations and entangling CZ gates (controlled-phase) rather than standard CNOT gates. Finally, the expectation values of the Pauli-Z operators are measured across all qubits and mapped to the final output classes through a second classical fully connected layer.

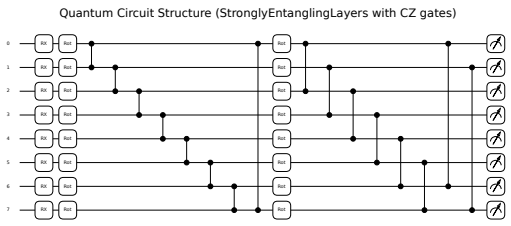

Figure 8: Model architecture of our backbone quantum circuit (with 8 qubits, 2 layers, and 3 rotation gates) using PennyLane's '*qml.StronglyEntanglingLayers*'.

Implementation details and hyperparameters for baselines and competing methods. We follow the overall setting in **Forward** (Patrini et al., 2017) to build the initial transition matrix. The key difference is that we train a two-layer shallow neural network on the training data without relying on anchor points. To ensure a fair comparison, we implement all baselines and competing methods using the same backbone QNN and optimizer with an identical batch size (128) and learning rate (0.01). Specifically, for **VolMinNet** (Li et al., 2021), we set the regularization coefficient to 10^{-3}. For **TVR** (Zhang et al., 2021), we set the regularization coefficient $\gamma = 0.1$ and the beta distribution parameters to $[0.999, 0.01]$. For **BLTM** (Yang et al., 2022), the upper bound threshold $\rho_{\max}$ is set to 0.6. For **CCR** (Cheng et al., 2022), the weight of the consistency regularization term is set to $\lambda = 0.3$. For competing methods validated on CheXpert, we conducted a grid search to fine-tune the key hyperparameters and, for the rest, followed the same settings as those in the original papers. For **DivideMix** (Li et al., 2020), we limited the warm-up period to 10 epochs to accommodate the

smaller dataset, while fixing the Beta parameter, probability threshold, and sharpening temperature all at 0.5. In **ELR** (Liu et al., 2020), we applied a regularization weight of 3.0 and set the temporal ensembling momentum to 0.7. The **VisualCheXbert** (Jain et al., 2021) was set up to utilize both label smoothing and KL-divergence. Regarding **BoMD** (Chen et al., 2023), we adopted a projection dimension of 128, maintained a memory bank queue size of 8192, and set the temperature parameter τ to 0.07. Finally, for **FixMatch** (Ihler et al., 2024), we established a high confidence threshold of 0.95, assigned an unlabeled loss weight of 1.0, and used an unlabeled data ratio of 5.

Table 4: Comparison with baselines across five datasets under 10% instance-dependent noise (IDN) by mean(std) of F1 (%) over five runs. The best results are highlighted in bold.

Models	Breast	Pneumonia	Retina	Derma	Blood
DISC	65.71 (2.11)	83.36 (2.42)	27.08 (2.27)	25.94 (1.73)	66.88 (1.08)
ProMix	66.34 (2.03)	84.19 (2.08)	28.21 (2.01)	26.83 (1.61)	67.57 (0.98)
InstanceGM	65.88 (1.94)	83.74 (1.97)	28.76 (2.12)	27.19 (1.56)	67.96 (1.03)
π-LR	66.02 (1.86)	83.87 (1.89)	28.94 (2.05)	27.37 (1.42)	68.06 (0.95)
Ours	**69.07 (2.32)**	**86.15 (1.72)**	**33.53 (2.80)**	**28.67 (1.50)**	**69.25 (0.95)**

Validation under instance-dependent noise (IDN). To validate the effectiveness of our method under IDN, we follow the benchmark setup (Ma et al., 2025), which in turn follows (Xia et al., 2020), to construct instance-dependent noise and compare our method with the following leading baselines in the 10% IDN setting. We finetune the hyperparameters on validation sets and choose the setting of 0.4 η, 0.5 del, and 20 pat. For fair comparison, we use the same backbone model and optimization hyperparameters. **DISC** (Li et al., 2023) dynamically selects reliable instances and corrects suspicious labels during training in an instance-specific manner. **ProMix** (Xiao et al., 2023) combats label noise by maximizing the utility of clean samples while reducing the harmful effect of corrupted ones. We set the threshold as 0.9. **InstanceGM** (Garg et al., 2023) models instance-dependent label corruption with graphical modeling. We set the threshold as 0.5, and $\alpha = 4$, $\lambda_u = 22$, $T = 0.5$. **π-LR** (He et al., 2025) models each sample's confusing probability and refines labels probabilistically to handle IDN. We set both elr and consistency ratio as 0.75, $\lambda = 0.3$. As shown in Tab. 4, the results verify that our method achieves robust performance under IDN settings and outperforms these baselines.

Validation on CheXpert, a real-world, noisy labeled dataset. We evaluate our proposed method on the real-world, noisy labeled chest radiograph dataset, CheXpert (Irvin et al., 2019), which consists of 224,316 images collected from 65,240 patients at Stanford Hospital. Labels for the training set were automatically extracted from radiology reports using a natural language processing (NLP) labeler, resulting in four categories: positive, negative, uncertain, and unmentioned. Following standard evaluation protocols, we focus on the five competition pathologies: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. For performance assessment, we utilize the official test set, comprising 500 studies annotated by the consensus of five board-certified radiologists, which serves as the clean ground truth. To simulate the data scarcity, we employ a stratified sampling strategy to sample 10% of the training set. Also, we apply the U-Zeros approach to assign the uncertain labels. We finetune the hyperparameters on validation sets and choose the setting of 0.4 η, 0.5 del, and 15 pat. In experiments, we compute the average AUC and F1 of the above five competition pathologies, and repeat experiments five times. The results are reported in mean (std). We compare our method with both generic noisy-label learning baselines and domain-specific medical imaging methods, which are widely applied on the CheXpert dataset. **DivideMix** (Li et al., 2020) is a semi-supervised framework that uses a Gaussian Mixture Model to dynamically divide samples into labeled (clean) and unlabeled (noisy) sets for co-training. **VisualCheXbert** (Jain et al., 2021) is a domain-specific method that utilizes a BERT-based text encoder

Table 5: Comparison of competing methods on real-world noisy labeled dataset, CheXpert. We report the mean AUC (%) and F1 score (%) $\pm$ standard deviation over five independent runs. The best results are highlighted in **bold**.

Methods	**AUC (%)**	**F1 (%)**
CE (U-Zero)	87.10 (0.15)	64.20 (0.05)
DivideMix	87.88 (0.22)	65.85 (0.15)
VisualCheXbert	88.11 (0.18)	66.22 (0.04)
ELR	87.52 (0.20)	65.08 (0.09)
BoMD	88.47 (0.12)	67.55 (0.03)
FixMatch	88.03 (0.25)	66.59 (0.16)
Ours	**89.85 (0.10)**	**68.94 (0.02)**

to extract high-quality soft labels from radiology reports to supervise the image classifier. **ELR** (Liu et al., 2020) is a regularization technique that leverages the "early learning" phenomenon to prevent deep networks from memorizing false labels by enforcing prediction consistency. **BoMD** (Chen et al., 2023) is a state-of-the-art medical method that employs text-driven multi-label descriptors to guide the image model in learning distinct pathological features under noise. **FixMatch** (Ihler et al., 2024) is a semi-supervised learning method that enforces consistency between the model's predictions on weakly-augmented and strongly-augmented versions of the same image.

Tab. 5 shows that our method achieves a new state-of-the-art performance on the CheXpert dataset, outperforming both generic noisy-label learning baselines and domain-specific medical imaging methods. Specifically, our method reaches an AUC of 89.85% and an F1 score of 68.94%, surpassing the previous strongest competitor, BoMD. It is worth noting that while general noisy-label methods like DivideMix and ELR provide improvements over Cross-Entropy baseline (87.88% and 87.52% vs. 87.10% in AUC), their gains are limited compared to our approach. This suggests that general noise-label learning strategies developed for natural images (e.g., CIFAR) may struggle under data scarcity. The results validate that our methods work effectively on real-world, noisy labeled datasets.

Analysis of trainable parameters. We compare the number of trainable parameters among the backbone models, SNN, DNN, and QNN. As shown in Tab. 6, QNN contains only 586 trainable parameters when $K = 2$ (the scale is normalized by QNN). In contrast, the shallow classical model contains 2,146 parameters, which is about $3.66\times$ larger than ours, while the deep classical model with 10 residual blocks contains 23,266 parameters, about $39.70\times$ larger. This substantial reduction in model size suggests that our quantum model provides a highly parameter-efficient alternative, achieving a compact architecture while retaining expressive modeling capacity.

Table 6: Model size analysis.

Model	$\lvert\boldsymbol{\theta}\rvert$	Scale
SNN	2,146	$3.66\times$
DNN	23,266	$39.70\times$
QNN	586	$1\times$

Analysis of smoothness in QNNs. To further examine how the smoothness affects QNN, we analyze the F1 scores of the variant "QNN+TS+SLT" under 50% uniform noise across different temperature values. In Fig. 9, we show a row-wise normalized heatmap, where each row corresponds to a dataset, and each column corresponds to a temperature value; within each row, darker cells indicate better F1 for that dataset. A clear and consistent trend emerges across all five datasets: our method achieves the best F1 at Temp = 1.0, namely, when no temperature scaling is applied to the QNN. This observation suggests that the intrinsic smoothness of QNNs already mitigates the adverse effects of overconfident predictions, making additional output-scale adjustment unnecessary. In turn, this provides further evidence for the advantage of using QNNs as the backbone model in our framework.

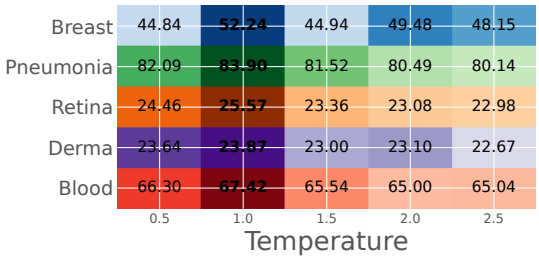

Figure 9: An Analysis of the impact of smoothness on QNNs across five datasets under 50% uniform noise. We apply temperature scaling to the output of QNNs before utilizing SLT to simulate different scales of smoothness and report the F1 score (%).

Analysis of update strategies for the transition matrix. Because frequent updates can destabilize the transition, we examine how different update strategies affect model performance on validation sets. We report results on the Breast dataset and present the remaining results in Fig. 11. As shown in Fig. 10, repeatedly executing the strategy ("Recur") clearly outperforms a single execution ("Once"). The combined strategy "NDU+PTU" is the most robust, yielding the best F1 with smaller variance across hyperparameters. Based on these observations, we adopt "NDU+PTU" and apply it recurrently during training. Overall, based on the analysis of η, del, and pat, we observe that while SLT

Table 8: Optimal hyperparameters, η, del, and pat, per dataset and noise type. 'UN', 'CF', and 'CM' denote the noise types of Uniform, Cyclic-flipping, and Custom-mapping, respectively. We adopt the update mode "NDU+PTU" for all datasets.

	UN			CF			CM		
	η	del	pat	η	del	pat	η	del	pat
Breast	0.6	0.3	20	0.9	0.3	20	0.9	0.3	20
Pneumonia	0.1	0.7	20	1.0	0.3	20	1.0	0.3	20
Retina	0.1	0.3	20	0.2	0.3	20	0.1	0.7	15
Derma	1.0	0.7	10	0.3	0.3	10	0.4	0.7	20
Blood	0.4	0.5	15	0.3	0.5	15	0.4	0.7	20

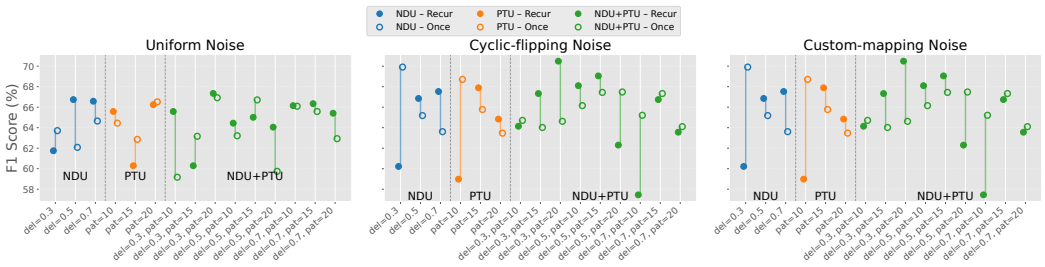

Figure 10: Analysis of update strategies (NDU, PTU, and NDU+PTU) under two scenarios (Recur and Once) for the transition matrix under three noise types (Uniform, Cyclic-flipping, and Custom-mapping) with a 10% noise ratio on Breast. "NDU" delays the transition matrix update until a specified fraction ("del") of total training iterations is completed. "PTU" triggers an update only after performance stagnates (e.g., no F1 increase or entropy decrease) for "pat" consecutive rounds. "NDU+PTU" signifies that both strategies are implemented concurrently. "Recur" and "Once" denote whether or not the above strategies are applied repeatedly whenever their conditions are met. The vertical line indicates the gap in the F1 score under two scenarios.

Table 7: Comparison with baselines (CE, Forward) and competing methods across Derma and Blood, under three noise settings (UN: Uniform, CF: Cyclic-flipping, CM: Custom-mapping) at noise levels 10%/30%/50%. We report mean(std) of F1 (%) over five runs.

Noise	Models	Derma			Blood		
		10%	30%	50%	10%	30%	50%
UN	CE	24.01 (1.25)	21.08 (2.32)	19.82 (2.98)	68.78 (0.77)	67.24 (0.61)	65.29 (0.58)
	Forward	27.13 (1.33)	25.93 (1.75)	22.94 (1.80)	69.03 (0.93)	68.56 (0.70)	66.74 (0.65)
	T-Revision	16.44 (3.11)	16.69 (4.37)	14.61 (3.83)	64.37 (5.97)	59.91 (6.67)	57.91 (6.68)
	Dual-T	27.28 (0.74)	25.03 (1.67)	22.75 (1.79)	69.13 (0.88)	63.78 (3.90)	59.15 (3.86)
	VolMinNet	26.09 (1.94)	24.89 (2.16)	21.94 (2.57)	68.47 (1.89)	62.41 (4.62)	57.23 (4.39)
	TVR	17.87 (2.75)	16.06 (4.47)	11.66 (3.67)	53.18 (8.08)	46.83 (8.99)	39.42 (9.56)
	BLTM	19.56 (1.77)	18.59 (1.57)	16.90 (3.21)	68.61 (1.38)	64.98 (2.61)	57.51 (3.65)
	CCR	25.32 (0.74)	25.06 (0.43)	22.86 (1.99)	68.62 (0.67)	67.87 (1.58)	66.82 (1.09)
	Ours	**29.24 (1.71)**	**26.95 (1.34)**	**23.87 (1.56)**	**69.88 (0.63)**	**69.14 (1.26)**	**67.42 (1.51)**
CF	CE	21.29 (2.34)	19.77 (2.02)	15.59 (3.86)	68.10 (0.89)	66.10 (1.15)	38.97 (4.12)
	Forward	26.56 (1.67)	24.37 (1.91)	15.90 (2.38)	70.17 (0.97)	68.49 (1.39)	46.38 (3.96)
	T-Revision	21.52 (3.48)	18.88 (4.02)	14.39 (4.57)	64.24 (5.18)	54.76 (4.49)	44.53 (4.10)
	Dual-T	26.66 (1.55)	24.75 (1.81)	15.40 (1.38)	70.14 (0.46)	67.63 (3.65)	50.74 (6.42)
	VolMinNet	26.96 (1.89)	23.73 (2.62)	14.87 (0.78)	69.57 (0.19)	68.27 (3.61)	48.98 (5.01)
	TVR	20.46 (4.13)	15.32 (2.78)	13.33 (4.34)	59.74 (8.08)	55.01 (6.58)	42.58 (5.72)
	BLTM	22.62 (2.73)	19.68 (2.19)	16.19 (2.15)	67.71 (0.85)	58.96 (7.47)	42.93 (5.24)
	CCR	25.00 (1.03)	23.94 (1.62)	16.82 (2.64)	69.09 (1.38)	68.30 (2.84)	48.41 (5.33)
	Ours	**27.69 (1.52)**	**25.89 (1.33)**	**18.04 (2.12)**	**71.81 (0.95)**	**69.52 (0.86)**	**51.69 (3.47)**
CM	CE	20.44 (2.58)	19.29 (2.40)	15.14 (3.97)	69.84 (0.89)	66.58 (0.77)	40.06 (2.24)
	Forward	26.06 (1.26)	22.72 (1.35)	16.84 (3.47)	69.92 (0.70)	69.37 (0.28)	49.80 (1.79)
	T-Revision	20.55 (2.70)	18.74 (3.71)	16.15 (3.16)	66.99 (3.23)	54.70 (1.43)	40.67 (4.41)
	Dual-T	25.06 (1.56)	21.72 (1.21)	15.93 (2.75)	69.03 (0.68)	65.02 (5.76)	39.02 (7.78)
	VolMinNet	25.21 (0.03)	20.94 (2.27)	14.34 (3.91)	69.65 (1.69)	63.96 (4.59)	40.78 (5.27)
	TVR	20.25 (3.98)	16.09 (4.36)	12.83 (4.40)	57.20 (4.07)	55.19 (5.14)	44.85 (6.52)
	BLTM	21.24 (1.33)	18.70 (2.78)	17.86 (1.67)	69.04 (0.98)	57.77 (7.11)	46.59 (6.18)
	CCR	24.86 (0.86)	23.10 (1.86)	17.25 (2.84)	69.67 (0.68)	68.27 (0.69)	50.34 (5.54)
	Ours	**27.51 (1.92)**	**23.57 (1.68)**	**18.12 (3.15)**	**70.56 (0.97)**	**69.78 (0.53)**	**51.30 (3.68)**

benefits from parameter tuning to achieve higher F1 scores, the performance degradation with suboptimal parameters is mild. This demonstrates the framework's robustness to variations in hyperparameter settings. We summarize the optimal hyperparameters in Tab. 8.

Discussion on failure cases in trade-off analysis. Loss-correction approaches depend heavily on an accurately estimated initial transition matrix. Specifically, this reliance necessitates that the

backbone model's predictions adhere to a closed-set assumption. In our experiments, we observed that employing an insufficient number of qubits leads to failure in constructing the initial transition matrix. It is important to note, however, that this is not a unique limitation inherent to our proposed framework.

Additional Comparison. We present the rest of the comparison results in Tab. 7.

Hardware and software. Our experimental environment consisted of an Ubuntu 22.04.3 LTS system powered by an Intel Xeon w5-3433 CPU and a 48GB NVIDIA RTX A6000 GPU. The software stack included Python 3.11.8, PyTorch 2.2.2, and Pennylane 0.35.1. The time required for the model training (single run) ranges from 25s (smallest dataset) to 582s (largest dataset).

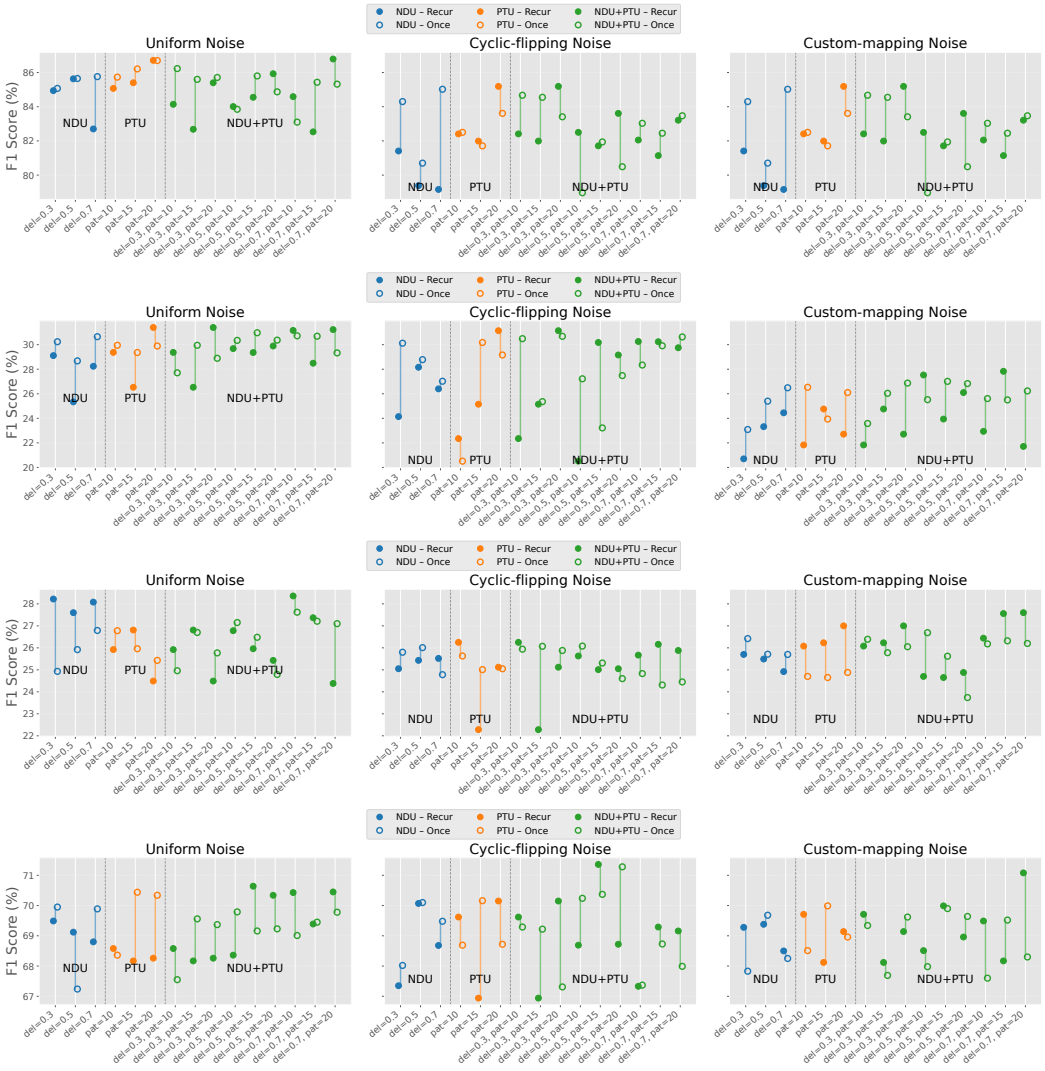

Figure 11: Analysis of update strategies (NDU, PTU, and NDU+PTU) under two scenarios (Recur and Once) for the transition matrix under three noise types (Uniform, Cyclic-flipping, and Custom-mapping) across four datasets in order (Pneumonia, Retina, Derma, and Blood). "NDU" delays the transition matrix update until a specified fraction ("del") of total training iterations is completed. "PTU" triggers an update only after performance stagnates (e.g., no F1 increase or entropy decrease) for "pat" consecutive rounds. "NDU+PTU" signifies that both strategies are implemented concurrently. "Recur" and "Once" denote whether or not the above strategies are applied repeatedly whenever their conditions are met. The vertical line indicates the gap in the F1 score under two scenarios.