WRITE YOUR OWN PROJECT TITLE HERE

RESEARCH ARTICLE

Author 1[1], Author 2[1], Author 2[2], YE Name[1], and M.R.C. Mahdy[1, *]

[1]Department of Electrical and Computer Engineering, North South University, Bashundhara, Dhaka
[2]Department of Electrical and Computer Engineering, University of Dhaka, Dhaka, Bangladesh

ABSTRACT

Very recently, studies have shown that quantum neural networks surpass classical neural networks in tasks like image classification when a similar number of learnable parameters are used. However, the development and optimization of quantum models are currently hindered by issues such as qubit instability and limited qubit availability, leading to error-prone systems with weak performance. On the other hand, classical models can exhibit high-performance owing to substantial resource availability. As a result, more studies have been focusing on hybrid classical-quantum integration to leverage the advantages of both paradigms. A line of research particularly focuses on transfer learning through classical-quantum integration or quantum-quantum approaches. Unlike previous studies, this paper introduces a new method to transfer knowledge from classical to quantum neural networks using knowledge distillation, effectively bridging the gap between classical machine learning and emergent quantum computing techniques. We adapt classical convolutional neural network (CNN) architectures like LeNet and AlexNet to serve as teacher networks, facilitating the training of student quantum models by sending supervisory signals during backpropagation through KL-divergence. The approach yields significant performance improvements for the quantum models by solely depending on classical CNNs, with quantum models achieving an average accuracy improvement of 0.80% on the MNIST dataset and 5.40% on the more complex FashionMNIST dataset. Applying this technique eliminates the cumbersome training of huge quantum models for transfer learning in resource-constrained settings and enables re-using existing pre-trained classical models to improve performance. Thus, this study paves the way for future research in quantum machine learning (QML) by positioning knowledge distillation as a core technique for advancing QML applications.

Keywords Parameterized quantum circuits · Quantum superposition · Knowledge distillation · Convolutional neural networks · Classification

1 Introduction

Quantum machine learning (QML) holds great promise to revolutionize computational paradigms by offering unprecedented computational speedups for certain tasks through parallel computation leveraging quantum superposition and entanglement. In particular, quantum neural networks (QNNs) have shown promising results compared to classical approaches by outperforming classical models in image classification tasks and converging faster when similar number of learnable parameters are used [1, 2]. A hybrid security system has been proposed in this study [3]. However, each of these approaches has its own limitations. Neural error mitigation requires a large amount of training data and computational resources to train the neural networks that correct the quantum errors[4]. Variational quantum-neural hybrid error mitigation introduces additional variational parameters that need to be optimized, which can increase the complexity and instability of the quantum optimization process. Resource-efficient quantum circuit design depends on the specific choice of the linear transformation and the quantum feature map, and may not be applicable to arbitrary

* Corresponding author. *E-mail address*: mahdy.chowdhury@northsouth.edu (M.R.C. Mahdy).

quantum tasks [5]. Quantum-to-quantum transfer learning assumes that the source and target quantum systems or tasks are sufficiently similar, and may not work well for heterogeneous or diverse quantum domains. Classical-to-quantum transfer learning relies on the classical neural network to extract meaningful features from the data, which may not capture the quantum correlations or entanglement that are essential for quantum advantages.

Transfer learning, especially using knowledge distillation [6], effectively navigates these challenges by offering a straightforward alternative where large pre-trained models share their learned knowledge with smaller ones in a teacher-student setup. This approach obviates the necessity of burdensome training and error mitigation to improve model performance by depending solely on the logits of large pre-trained models, assuaging the computational cost. Furthermore, this method also circumvents the limitations of the available resources by enabling smaller models to inherit robust, pre-optimized decision boundaries and representational features from the teacher models. KD can be used to transfer the knowledge learned by a large or complex neural network (the teacher) to a smaller or simpler neural network (the student). It can also reduce the noise sensitivity of the student network, and enable cross-domain or cross-hardware applications. In addition, it can enhance the interpretability and explainability of the student neural network, by revealing the underlying physical rules or symmetries that govern the performance enhancement. As a result, knowledge distillation can be used and applied as a promising technique for neural network design and optimization.

Knowledge distillation in neural networks can be accomplished in various ways: **(1)** Response-based knowledge distillation [7]: This approach transfers the knowledge from the teacher network to the student network by matching their outputs or responses on a given dataset. The student network learns to mimic the teacher networks predictions, probabilities, or logits, and thus inherits the teacher networks generalization ability. **(2)** Feature-based knowledge distillation [7]: This approach transfers the knowledge from the teacher network to the student network by matching their intermediate features or representations. The student network learns to extract similar features as the teacher network, and thus captures the teacher networks discriminative power. **(3)** Relation-based knowledge distillation [7]: This approach transfers the knowledge from the teacher network to the student network by matching their relations or interactions among the inputs, outputs, or features. The student network learns to preserve the same relations as the teacher network, and thus acquires the teacher networks structural or semantic information.

Recent studies have explored transfer learning in quantum neural networks with response-based knowledge distillation using a Quantum-to-quantum approach [4]. A large domain of Knowledge Distillation remains unexplored: Classical-to-quantum knowledge transfer. Given the constraints of quantum machine learning, it is crucial to explore the possibilities that a quantum model can enhance its performance solely based on the outputs of classical neural networks. This leads us to the research questions of this study, based on which we design the methodology, experiment rigorously, analyze and conclude our findings in this paper:

- **Research Question 1:** Can we transfer the knowledge from a classical neural network (the teacher) to a quantum neural network (the student) using knowledge distillation?
- **Research Question 2:** How does knowledge distillation from classical neural network to quantum neural network affect the performance of the quantum model?

In this paper, we show theoretically and experimentally that knowledge transfer from classical to quantum models using knowledge distillation indeed improves the performance of the quantum models drastically. This method obviates the dependency of QNNs upon larger QNNs, thus eliminating the necessity of hectic model training, quantum error-mitigation and resource handling for improving the performance of quantum models. To demonstrate our claim, we conduct extensive experiments on two widely used datasets: MNIST [8] and FashionMNIST [9]. We use multiple classical convolutional neural networks as teacher models to train on these datasets and generate the outputs for use by the quantum student models. We also use multiple quantum student models like 3-qubit and 4-qubit variational quantum circuits, to learn from the outputs of the classical teacher models and improve performance. We compare the performance of the quantum student models with and without knowledge distillation, and show that knowledge distillation can significantly improve the accuracy, robustness, and efficiency of the quantum models. Moreover, we also use ensembles of the classical teacher models, and investigate how the performance of the quantum student models varies with the number of teacher models used. We find that using more teacher models can provide more diverse and complementary information for the quantum student models, and thus further enhance the knowledge transfer effect.

The methodology followed in this study is visualized in Figure 1. The figure shows that after following data preprocessing steps for classical and quantum models respectively, the classical data is passed through a classical network containing teacher CNN model(s), whereas the quantum data is passed through student QNN model(s). After that, using the outputs of the models, the KL-Divergence is calculated and added to the loss of the student model to calculate a total loss which is used to improve performance of the models. Adding the KL-divergence acts as a supervisory signal while updating the parameters of the quantum model. In this teacher-student setup, the student QNN thus tries

to mimic the outputs of the teacher CNN and knowledge transfer is accomplished from classical to quantum models solely based on the mimicry.

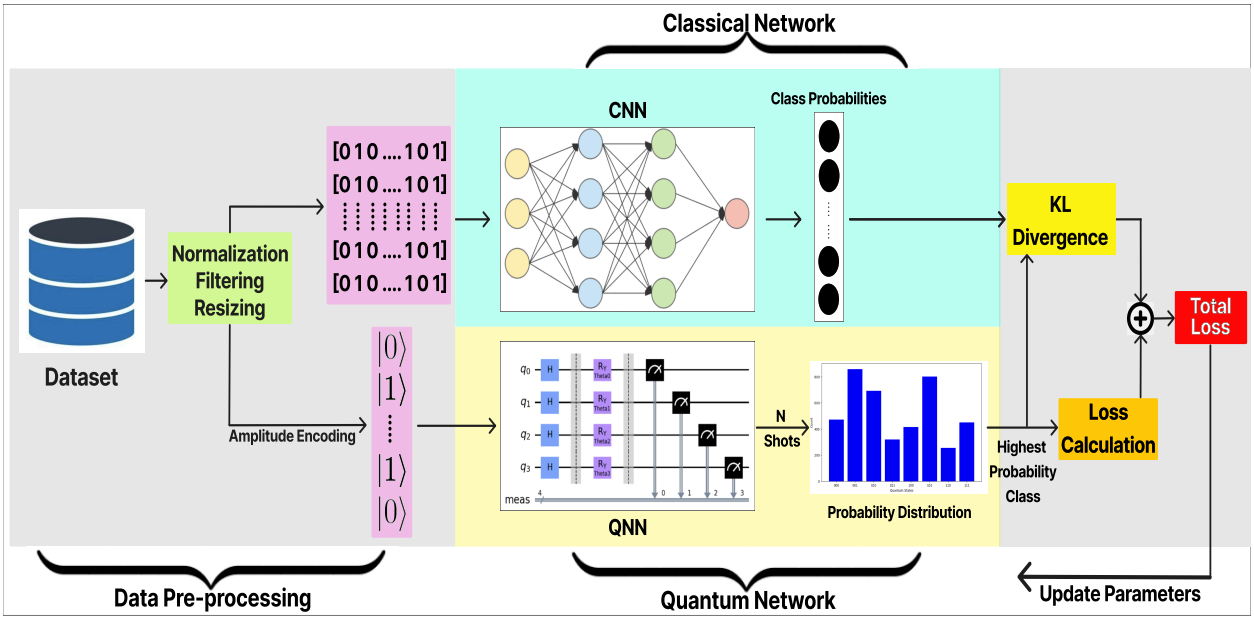


Figure 1: A visual representation of the research methodology followed in this research. At the very beginning, data pre-processing is carried out individually for classical and quantum model training. Following that, the data is passed through the respective models. A pre-trained CNN is used in the classical network. The outputs from the CNN and the QNN are used to calculate the KL-divergence and add that to the loss of the quantum network to aid in backpropagation.

Contributions. The contributions of this research can be summarized as follows:

- We propose a novel hybrid knowledge distillation framework to transfer the knowledge from multiple classical teacher models to quantum student models, which establishes the re-usability of existing pre-trained classical models in quantum training and eliminates the necessity of cumbersome QNN training.
- We systematically investigate the effect of knowledge distillation from classical neural networks to quantum neural networks on the performance of the quantum models.
- We conduct extensive experiments on two widely used datasets, MNIST and FashionMNIST, and compare the performance of various quantum student models with and without knowledge distillation.
- sjsahsdkj

The rest of the paper is organized as follows: Section 2 discusses the related works, providing context and background for our approach. Section 3 describes the methodology, detailing the mathematical foundations and our approach to knowledge distillation. In Section 4, we present the experiments and results, along with the datasets used, models utilized and the specifics of our implementation. Section 5 presents the findings of this study and dives into their in-depth discussion, interpreting the results and their implications. Section 6 concludes the paper and suggests directions for future work, highlighting the potential applications and extensions of our research. Finally, we acknowledge the contributions of the people that have assisted in this research. A new section is **??**

2 Related Works

2.1 Quantum Machine Learning: Problems and Challenges

Quantum Machine Learning (QML) aspires to harness the properties of quantum mechanics to revolutionize computation, but the path is restricted with various challenges and obstacles. The concept of quantum supremacy, where quantum systems perform tasks beyond the reach of classical computers, has seen experimental strides as evidenced by [10] and [11]. However, these pioneering steps also underscore the fragility of quantum states and their vulnerability

to decoherence, a problem that Shors early work [12] sought to address through quantum error correction, a technique still pivotal in contemporary research [13]. The transient stability of qubits, a topic explored by [14], and the intricate challenge of manipulating quantum states, as demonstrated by [15], add layers of complexity to quantum computing.

The Noisy Intermediate-Scale Quantum (NISQ) era, defined by the presence of noisy quantum bits, has constrained the scalability of quantum computers, a concern that [16] aimed to tackle through the quantum Boltzmann machine, an approach designed to work within these noisy environments. Concurrently, the development of quantum algorithms and neural networks that can operate under such noise is the focus of [17] and [18], who seek methods to optimize QML models despite these limitations. Data encoding poses a unique challenge in QML, as most algorithms are tailored for inherently quantum data. [19] delve into creating quantum-enhanced feature spaces, while [20] look into reinforcement learning agents that can operate within the quantum realm. These studies offer insights into the potential application domains for QML, suggesting that despite current limitations, certain tasks may still be within reach. Furthermore, the practical implementation of quantum computing and the exploration of its utility across various domains have been the subject of investigation by [21], who discuss the variational quantum algorithms that could be the cornerstone of future quantum applications.

Despite the under-developed stage of QML, there is an ongoing effort to address its challenges. [22] provides a comprehensive analysis of the prospects and barriers that lie ahead, ensuring that the community remains grounded in reality while exploring the vast potential of quantum technologies. The integration of efforts from all these studies manifests a commitment to overcoming the current hurdles, indicating a future where quantum and classical computing may converge to solve problems that were once thought unsolvable. In this research, we commit to this challenge and address it using a hybrid classical-quantum architecture and demonstrate that the methodology takes handling the challenges faced in the realm of QML one step further.

2.2 Error Mitigation and Resource Handling in Quantum Computing

The pursuit of reliable quantum computing is marked by a need to address the fundamental limits of quantum error mitigation. Works such as those by [23] provide a foundational understanding of the theoretical bounds in quantum systems, crucial for error correction techniques in neural networks. This is further explored through the lens of scalability and statistics of errors in recent advancements within noisy intermediate-scale quantum (NISQ) technologies [24]. Reference-state error mitigation strategies [25] have shown promise in improving the accuracy of quantum computations, which is vital for the fidelity of knowledge distillation processes. Moreover, the extension of computational reach through error mitigation in noisy quantum processors [26] has direct implications for the robust execution of quantum neural networks. The integration of learning-based techniques for quantum error mitigation [27] parallels the adaptive nature of knowledge distillation, suggesting a methodological symmetry that could be harnessed for transferring knowledge. The evaluation and benchmarking of error mitigation techniques are critical for determining the most effective methods for application in quantum neural networks [28]. This selection is further informed by the development of a unified approach to data-driven error mitigation [29], offering a pathway to harnessing empirical data in the error correction process. Techniques for mitigating errors in quantum approximate optimization [30] can also inform the optimization strategies within the knowledge distillation framework.

On the front of resource handling, the importance of optimizing resource efficiencies is underscored by [31], drawing attention to the balance between computational capabilities and resource constraints. This optimization is directly applicable to the development of quantum neural networks, where the allocation and management of quantum resources are crucial. Furthermore, the critical challenges of scaling quantum computation, as outlined by [32], must be addressed to enable the practical implementation of fault-tolerant quantum neural networks. Insights into the cross-discipline theoretical research on quantum computing, such as those presented by [33], highlight the urgency and significance of resource management in the context of knowledge distillation. Additionally, the potential applications of quantum computing in specific domains like renewable energy optimization [34] illustrate the diverse utility and the resource handling strategies that may be required for energy-efficient quantum neural network operations. Differently from previous approaches, we completely nullify the need for QNN optimization using error mitigation or resource handling. Rather, keeping the student quantum model small and static, we depend fully on pre-trained classical models and their ensembles to improve the performance of the QNN, thus addressing the challenges of QML in a new, different way.

2.3 Transfer Learning in Quantum Machine Learning

Transfer learning has emerged as a potent strategy in classical machine learning, enabling models trained on one task to be repurposed for another related task [35, 36]. In quantum machine learning (QML), this approach promises to address scalability and adaptability challenges that are crucial in the practical application of neural-network quantum

states (NNQS) [35]. Scalability, a primary concern in QML, is particularly relevant for tasks that require models to extrapolate learned behaviors to larger quantum systems. Protocols inspired by physics have shown that features learned from smaller systems can significantly accelerate and refine the learning process when applied to more complex systems, as evidenced in the one-dimensional transverse field Ising and Heisenberg XXZ models [35]. The versatility of transfer learning is further showcased by its application to the optimization of neural network potentials, exemplified by ANI-1x, which was retrained to near gold-standard quantum mechanical accuracy. This highlights transfer learning's broad potential across disciplines, including materials science, biology, and chemistry [36, 37].

Integrating classical and quantum neural networks has extended transfer learning protocols into classical-quantum hybrid domains. By combining pre-trained classical networks with variational quantum circuits, researchers have opened up new possibilities for processing high-dimensional data such as images before entangling them within the quantum processing framework [38]. This hybrid approach not only utilizes the strengths of classical networks in data handling but also harnesses the quantum advantages for feature processing, potentially revolutionizing fields like image recognition and quantum state classification [38, 39]. Moreover, the concept of knowledge distillation has been introduced into the realm of QML, allowing the creation of new quantum neural networks (QNNs) through approximate synthesis [4]. This methodology enables the reduction of circuit layers or the alteration of gate sets without the necessity of training from scratch. Empirical analyses suggest that significant reductions in circuit complexity can be achieved while simultaneously improving accuracy, even under noisy conditions, underscoring the effectiveness of transfer learning techniques in the development of QNNs [4].

In terms of practical applications, transfer learning has been successfully applied to improve the accuracy and efficiency of surrogate models used in design optimization [40], and disease detection [41, 42], demonstrating its efficacy in enhancing computational time and model performance. Furthermore, leveraging pre-existing computational datasets through deep transfer learning constructs robust prediction models that outperform those based solely on theoretical calculations [43]. The domain of spoken command recognition (SCR) has also benefitted from the implementation of hybrid transfer learning algorithms. By transferring knowledge from a pre-trained classical network to a hybrid quantum-classical model, researchers have been able to significantly improve performance on SCR tasks, showcasing the practical benefits of transfer learning in QML [44].

Quantum Convolutional Neural Networks (QCNNs) are gaining popularity in both quantum and classical data classification recently [45]. The adoption of transfer learning in QML is proposed to circumvent the scalability constraints of quantum circuits in the noisy intermediate-scale quantum (NISQ) era. For instance, research has shown that small QCNNs can effectively solve complex classification tasks by leveraging pre-trained classical CNNs, thus bypassing the need for expansive quantum circuitry [44]. This approach is validated through numerical simulations of QCNNs, which demonstrate a superior classification accuracy over classical models when pre-trained on different datasets. The synergy between classical and quantum networks is evident when a classical CNN, trained on FashionMNIST data, is utilized to enhance the performance of a QCNN for MNIST data classification [45]. The empirical results advocate for transfer learning from classical to quantum CNNs, revealing a performance that significantly outshines classical transfer learning models under analogous training conditions. These findings not only elucidate the potential of knowledge transfer to optimize the utilization of QCNNs in the NISQ era but also underscore the practicality of hybrid classical-quantum models in overcoming the limitations posed by the present-day quantum technologies. Driven by these findings, our approach is designed to transfer knowledge from classical CNNs to quantum models. However, differently from [44] and [45], we use knowledge distillation and show empirically that knowledge transfer is possible across homogeneous datasets in quantum models by simply trying to mimic the soft targets of classical CNNs instead of hectic training by freezing the pre-trained model and transferring the knowledge to a QNN during training in the previous approaches.

3 Methodology

3.1 Problem Formulation and Integration of Knowledge Distillation

So here is our methodology. The methodology diagram is shown in figure 2.

$$|B_{12}\rangle_{cd} = \frac{2}{\sqrt{2}} \left(|p\rangle \otimes |m\rangle + (-1)^n |1\rangle \otimes |1 \oplus m\rangle \right) \tag{1}$$

We consider a classical neural network as the teacher model and a quantum neural network (QNN) as the student model. The teacher model is denoted by $f_t : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. Similarly, the student model is denoted by $f_s : \mathcal{X} \to \mathcal{Y}$.

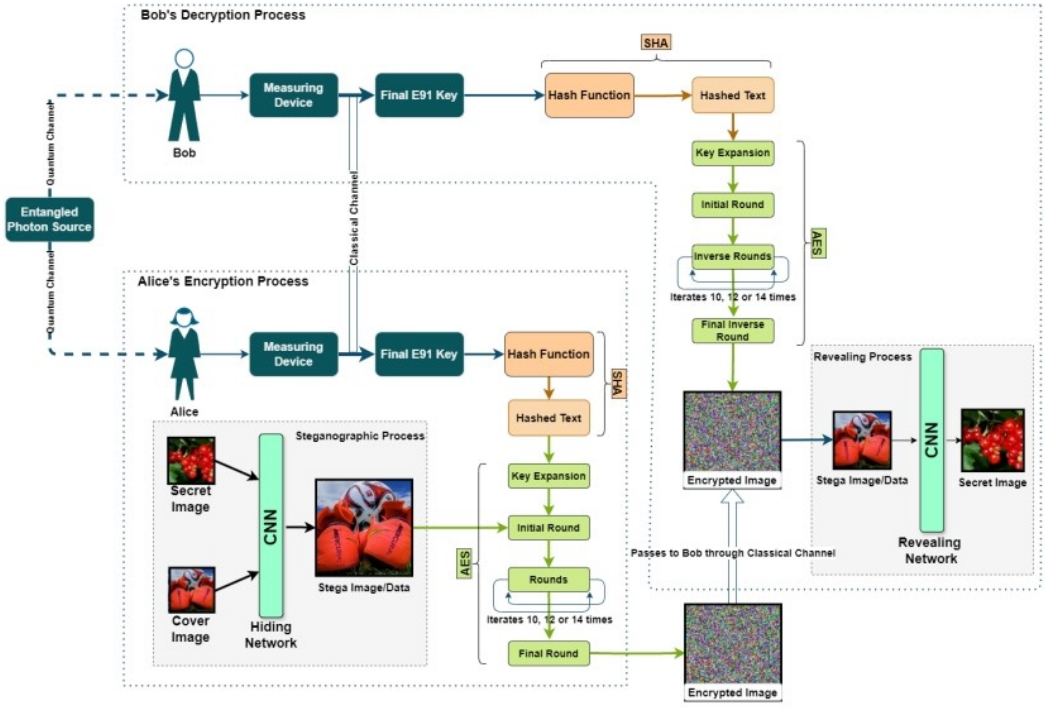


Figure 2: Methodology diagram

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, the task is to train the student QNN to approximate the mapping of the teacher model. The performance of the QNN can be evaluated by a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, which measures the discrepancy between the predictions of the QNN and the ground truth labels. The objective function for the QNN is given in Equation 2 as:

$$\mathcal{O}(f_s) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_s(\mathbf{x}_i), y_i). \tag{2}$$

The loss of the model outputs, $\mathcal{L}$, is calculated using cross-entropy loss [46] according to Equation 3:

$$\mathcal{L}_{\text{cross-entropy}}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3}$$

where C is the total number of classes, y_i is the ground truth label indicating the probability of the predicted class, and $\hat{y}_i$ is the predicted probability that the observation is of class i.

Knowledge Distillation (KD) aims to transfer knowledge from the teacher model to the student model by minimizing a distillation loss $\mathcal{L}_{KD}$. The KD process involves a soft target τ which controls the smoothness of the teachers output probability distribution. The distillation loss can be defined using the Kullback-Leibler (KL) divergence [47] in Equation 4 as:

$$\mathcal{L}_{KD}(f_s, f_t) = \frac{1}{N} \sum_{i=1}^{N} KL\Big(\sigma\big(\frac{f_t(\mathbf{x}_i)}{\tau}\big) \,||\, \sigma\big(\frac{f_s(\mathbf{x}_i)}{\tau}\big)\Big), \tag{4}$$

where σ denotes the softmax function and KL denotes the Kullback-Leibler divergence. The softmax function is given according to Equation 5 by:

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{5}$$

where $\sigma(\mathbf{z})_i$ is the output of the softmax function for the i-th class, z_i is the i-th element of the input vector $\mathbf{z}$, and K is the total number of classes. The exponential function e^{z_i} ensures that all outputs are non-negative and the denominator normalizes the values to ensure they sum up to 1, thus forming a valid probability distribution.

And, the KL-divergence is calculated as per Equation 6:

$$KL(P||Q) = \sum_{i=1}^{K} P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{6}$$

where P and Q are discrete probability distributions. $P(i)$ represents the probability of the i-th class according to the true label distribution, typically obtained from the softmax function of the teacher model $\sigma(f_t(\mathbf{x}_i))$, and $Q(i)$ is the probability of the i-th class according to the predicted label distribution from the student model $\sigma(f_s(\mathbf{x}_i))$. K is the total number of classes. The KL divergence measures the difference between two probability distributions over the same variable. The total loss for the student QNN, incorporating both the original objective and the distillation loss, is given in Equation 7 by:

$$\mathcal{O}_{total}(f_s) = \mathcal{O}(f_s) + \lambda \mathcal{L}_{KD}(f_s, f_t), \tag{7}$$

where λ is a hyperparameter that balances the importance of the original loss and the distillation loss.

The student QNN is trained by optimizing $\mathcal{O}_{total}$ through a quantum optimization algorithm. The goal of the optimization is to minimize the loss, which can be formulated in Equation 8 as:

$$f_s^* = \arg\min_{f_s} \mathcal{O}_{total}(f_s). \tag{8}$$

3.2 Quantum Superposition and the Mathematical Intuition Behind QNNs

A QNN can be represented by a parameterized quantum circuit (PQC) $U(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the vector of parameters. The output of the QNN for an input state $|\psi(\mathbf{x})\rangle$ prepared by a quantum feature map is given in Equation 9 by:

$$f_s(\mathbf{x}) = \langle \psi(\mathbf{x}) | U^\dagger(\boldsymbol{\theta}) O U(\boldsymbol{\theta}) | \psi(\mathbf{x}) \rangle, \tag{9}$$

where O is an observable corresponding to the output measurement.

The postulates of quantum mechanics [48] allow us to describe the state of a quantum system using a wave function or state vector in a Hilbert space. A quantum system in a state $|\psi\rangle$ can exist in a superposition of different basis states $\{|i\rangle\}$. This superposition is mathematically represented in Equation 10 as:

$$|\psi\rangle = \sum_i c_i |i\rangle, \tag{10}$$

where c_i are complex coefficients satisfying the normalization condition $\sum_i |c_i|^2 = 1$.

The probability $P(i)$ of the system being found in a particular basis state $|i\rangle$ upon measurement is given by the square of the modulus of the coefficient corresponding to that state: $P(i) = |c_i|^2$

The expectation value $\langle O \rangle$ of an observable O, which corresponds to a physical quantity that can be measured, is computed according to Equation 11 as:

$$\langle O \rangle = \langle \psi | O | \psi \rangle = \sum_{i,j} c_i^* c_j \langle i | O | j \rangle. \tag{11}$$

For an observable O with eigenstates $|o_k\rangle$ and eigenvalues o_k, the expectation value takes the form of Equation 12:

$$\langle O \rangle = \sum_k o_k P(o_k), \tag{12}$$

where $P(o_k)$ is the probability of obtaining the eigenvalue o_k upon measuring the observable O.

In the context of quantum computing, particularly in quantum neural networks, the output of the network can be associated with the expectation value of a certain observable. By designing the observable suitably, the expectation values can be interpreted as probabilities, providing us with a powerful tool to map quantum computations to classical outputs useful for tasks such as classification. This probabilistic interpretation is the key to the integration of quantum systems with machine learning, as it allows us to use quantum computations to produce outputs that can be understood and utilized within the classical framework of neural networks.

The knowledge transfer from the classical teacher model to the quantum student model is achieved by the optimization of the quantum circuit parameters $\boldsymbol{\theta}$ to minimize $\mathcal{O}_{total}$. This optimization can be performed using gradient-based methods where the gradient can be estimated via the parameter shift rule or other quantum backpropagation techniques.

3.3 Quantum Gradient Descent and Parameter Optimization

In the training of quantum neural networks (QNNs), the optimization of parameters $\boldsymbol{\theta}$ is crucial. Due to the nature of quantum computing, traditional backpropagation as used in classical neural networks is not directly applicable. Instead, we utilize the parameter shift rule to compute gradients of quantum circuits.

3.3.1 Parameter Shift Rule

The parameter shift rule is a method to compute the gradient of an expectation value of an observable with respect to the parameters of a quantum circuit [49]. For a parameter θ_j and an observable O, the gradient can be calculated using the following Equation 13 as:

$$\frac{\partial \langle O \rangle}{\partial \theta_j} = \frac{1}{2} \left[\langle O \rangle_{\theta_j + \frac{\pi}{2}} - \langle O \rangle_{\theta_j - \frac{\pi}{2}} \right], \tag{13}$$

where $\langle O \rangle_{\theta_j \pm \frac{\pi}{2}}$ denotes the expectation value of O when the parameter θ_j is shifted by $\pm \frac{\pi}{2}$ respectively.

3.3.2 Quantum Gradient Descent

Once the gradients are computed, the parameters can be updated using a gradient descent algorithm. The update rule at each iteration t is given by Equation 14:

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \eta \frac{\partial \mathcal{O}_{total}}{\partial \theta_j}, \tag{14}$$

where η is the learning rate.

The learning rate can be fixed or adaptively changed according to specific strategies, such as learning rate annealing [50] or using advanced optimizers like Adam [51], RMSprop [52], or AdaGrad [53], which are adapted for the quantum domain.

3.3.3 Quantum Circuit Learning

The training process involves iteratively adjusting the parameters $\boldsymbol{\theta}$ to minimize the loss function $\mathcal{O}_{total}$. This is referred to as quantum circuit learning and is given by Equation 8. The optimization process leverages the quantum superposition and entanglement to explore the parameter space more efficiently than classical methods. The inherent probabilistic nature of quantum measurements is accounted for in the optimization loop, making the training resilient to the stochastic nature of quantum mechanics. By integrating these quantum-specific optimization techniques, we aim to exploit the full potential of quantum neural networks while addressing the challenges posed by the hardware limitations and the unique aspects of quantum computation.

3.4 Amplitude Encoding in Quantum Neural Networks

Amplitude encoding [54] is a technique used in quantum computing to map classical data into quantum states. This method leverages the ability of quantum states to exist in superpositions, allowing an exponential compression of data compared to classical representations.

A classical vector $\mathbf{x} \in \mathbb{R}^n$ can be normalized and encoded into the amplitudes of a quantum state $|\psi\rangle$ in a Hilbert space of $\log_2(n)$ qubits, given according to Equation 15:

$$|\psi\rangle = \sum_{i=0}^{n-1} x_i |i\rangle, \tag{15}$$

where $\{x_i\}$ are the normalized values of the classical vector $\mathbf{x}$, and $\{|i\rangle\}$ represents the computational basis states.

The normalization ensures that $\sum_{i=0}^{n-1} |x_i|^2 = 1$, which is a requirement for any valid quantum state.

Amplitude encoding is necessary for efficiently utilizing the limited number of qubits available in current quantum hardware. It allows the encoding of a classical dataset with 2^n features into a quantum system with only n qubits, thus exponentially reducing the dimensionality of the data representation.

In practice, amplitude encoding can be implemented using various quantum gates and operations. For instance, the non-linear transformation given by $x' = \pi \cdot \tanh(x)$ can be applied to preprocess the data before encoding it into the amplitudes. Here, x' denotes the transformed feature that is to be encoded into the quantum state. The hyperbolic tangent function ensures that the features are scaled to the range $[-\pi, \pi]$, which is suitable for encoding into quantum state amplitudes. Then, we can map these preprocessed classical features into the amplitudes of a quantum state according to Equation 16:

$$|\psi(\mathbf{x})\rangle = \frac{1}{\|\mathbf{x}'\|} \sum_{i=0}^{n-1} x'_i |i\rangle, \tag{16}$$

where $\|\mathbf{x}'\|$ is the L2 norm of the vector $\mathbf{x}'$.

The encoding process translates into the preparation of quantum states through a sequence of quantum gates that transform the initial state $|0\rangle^{\otimes n}$ into the desired superposition state $|\psi(\mathbf{x})\rangle$. The specific sequence of gates depends on the quantum hardware and the form of the feature vector. In quantum neural networks, amplitude encoding serves as the crucial first step that translates classical information into a form that quantum layers can process, leading to the exploitation of quantum properties such as superposition and entanglement in subsequent computations. By using amplitude encoding, we ensure that classical data is compatible with quantum processing, allowing the QNN to leverage the advantages of quantum computation in solving machine learning tasks.

4 Experiments and Results

4.1 Datasets Used

In our experiments, we have employed two well-known datasets: MNIST and FashionMNIST. The datasets are shown in Figure 3. Two datasets are used to demonstrate the generalizability and broader application of the methodology in this study.

4.1.1 MNIST

The MNIST dataset [8] is a large database of handwritten digits that is commonly used for training various image processing systems. It contains 60,000 training images and 10,000 testing images, each of which is a grayscale image of size 28×28 pixels. The dataset is designed to allow researchers to benchmark their algorithms in classifying and recognizing handwritten numerical digits from 0 to 9. Due to its simplicity and the fact that it has been extensively studied, it serves as a standard for evaluating the performance of various machine learning techniques.

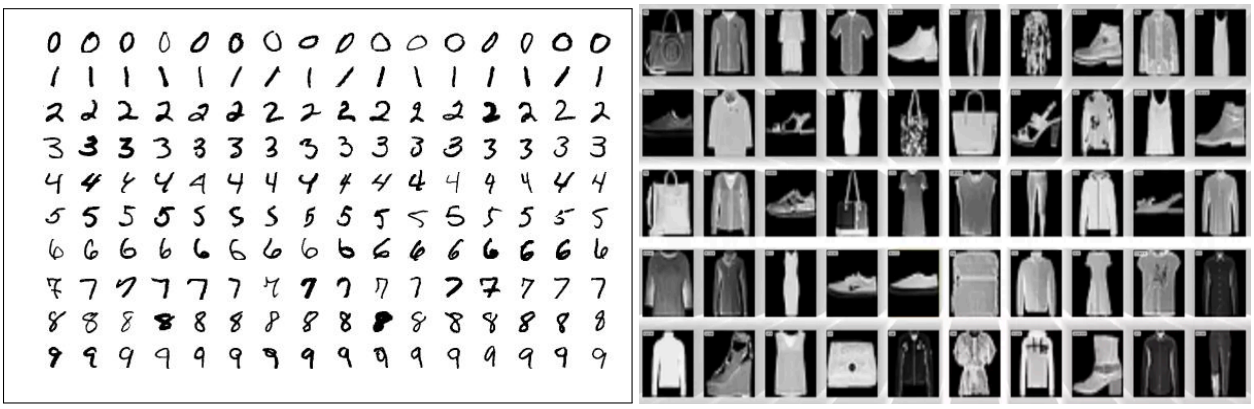
4.1.2 FashionMNIST

The FashionMNIST dataset [9] is a more recent dataset comprising 60,000 training samples and 10,000 test samples. Each sample is a 28×28 grayscale image, associated with a label from 10 classes. The classes represent different types of clothing and fashion items, such as T-shirts/tops, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. This dataset is intended to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms, with the added complexity of distinguishing between different types of clothing. Both datasets are preprocessed in a similar manner, normalized, and used as the basis for training and testing our models. By utilizing these datasets, we aim to show that our approach is not limited to a specific type of data or domain but can be applied to various image recognition tasks, highlighting its flexibility and wide applicability.

4.2 Dataset Filtration

In line with previous research that has addressed similar challenges in quantum machine learning [55, 4, 45], we filtered the MNIST and FashionMNIST datasets to reduce simulation time and focus on specific aspects of classification tasks. We created two subsets: MNIST 0-5, containing only the classes labeled from 0 to 5, and FashionMNIST 0-7, including classes from 0 to 7. Furthermore, within each class, we have randomly selected 150 samples to form the filtered datasets.

The filtration of datasets serves several purposes in our study. Firstly, it allows us to evaluate our method's effectiveness across a range of classes while keeping computational resources manageable. Quantum training can be inherently slow and resource-intensive due to the current limitations of quantum hardware and the complex nature of quantum computations. By reducing the number of classes and samples, we aim to demonstrate the feasibility of training



(a) (b)

Figure 3: The datasets used in the experiments. (a) The MNIST dataset of written digits having 10 classes. (b) The FashionMNIST dataset of clothing containing 10 classes.

quantum models without the need for extensive resources, which is crucial for the practical adoption of quantum machine learning.

Choosing a smaller subset of samples is motivated by the need for computational efficiency. Quantum computers and simulators, particularly those accessible for research purposes, are subject to resource constraints such as limited qubits and short coherence times. These constraints make training on full datasets impractical. The selection of 150 samples per class allows for a diverse enough representation to train on and validate our models, ensuring that the performance metrics reflect the models' ability to generalize from a limited sample size.

4.3 Classical Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have been paramount in the field of image recognition and classification. For the MNIST and FashionMNIST datasets, two architectures stand out due to their historical significance and performance: LeNet [56] and AlexNet [57]. In our experiments, we have used these two models individually and their ensemble by averaging the outputs of the final layers as classical teacher models. Both LeNet and AlexNet were originally intended for higher dimensional images. To accommodate the 28×28 pixel grayscale images of MNIST and FashionMNIST, we have modified the input layers of these models to accept single-channel 28×28 pixel input instead of their original input dimensions.

4.3.1 LeNet

LeNet is one of the earliest convolutional neural networks that significantly impacted the field of deep learning. It is a much simpler network compared to AlexNet, consisting of two convolutional layers followed by two fully connected layers. LeNet is particularly well-suited for handwritten digit recognition tasks like MNIST due to its architecture being designed for this purpose. In our experiments, we have used the original LeNet architecture by changing the input layer to accept images of dimension 28×28 instead of its original input image dimension of 32×32, as shown in Figure 4. The intermediate layers were constructed according to the original architecture with no change.

4.3.2 AlexNet

AlexNet is a deep CNN that marked a breakthrough in the ImageNet competition. It consists of five convolutional layers followed by three fully connected layers. AlexNet takes an RGB image of size 256×256 as input and uses 11×11 filters with a stride of 4 in the first convolutional layer. Since MNIST and FashionMNIST images are grayscale images of size 28×28, the architecture of AlexNet was modified in the intermediate layers. Figure 5 shows the followed architecture of AlexNet that was used instead: The modified architecture features an initial convolutional layer with 64 filters of size 3×3, stride 1, and padding 1, followed by ReLU activation and a max pooling layer with a 2×2 kernel and stride 2. The second layer has 192 filters of size 3×3 with padding 1, followed by ReLU activation and another 2×2 max pooling. The third convolutional layer consists of 384 filters with a kernel size of 3×3 and padding of 1, followed by ReLU activation. This is followed by a series of fully connected layers: the first with 4096 neurons, the second also with 4096 neurons, and the final layer corresponding to the number of classes. Dropout is applied before

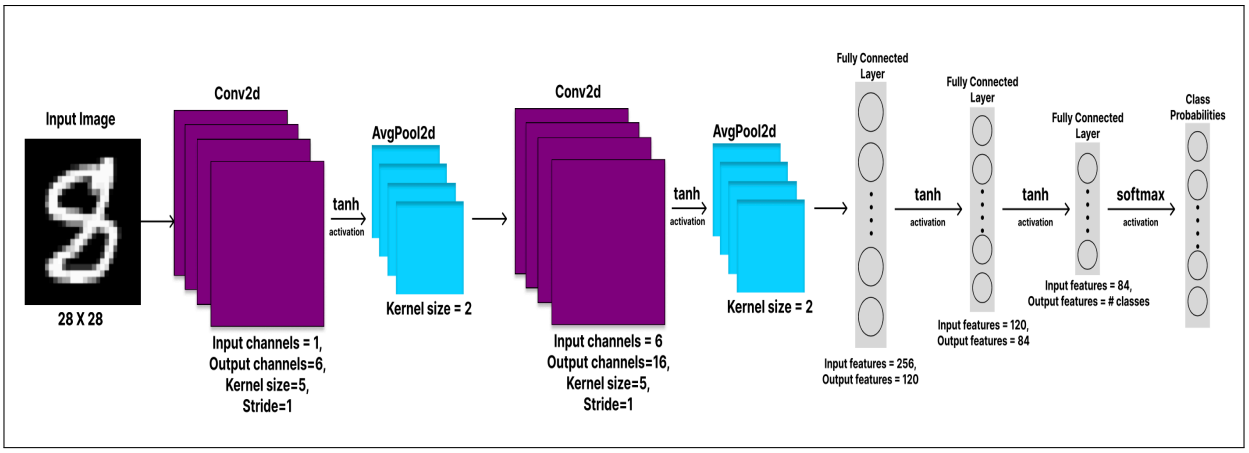


Figure 4: The modified LeNet architecture used in the implementation of the experiments. Two convolution layers, each followed by tanh activation and average pool layers were used. Finally, three fully connected layers were used to classify the images and get output probabilities from the model.

the first and second fully connected layers to prevent overfitting. The model is adapted for single-channel grayscale images and the smaller spatial dimensions of the MNIST and FashionMNIST datasets. The use of ReLU activation functions, dropout, and overlapping pooling makes AlexNet robust against overfitting and capable of learning complex patterns in image data. In the context of MNIST and FashionMNIST, AlexNet's architecture, although more complex than necessary for such simple datasets, provides an excellent baseline for performance due to its depth and capacity to learn intricate features.

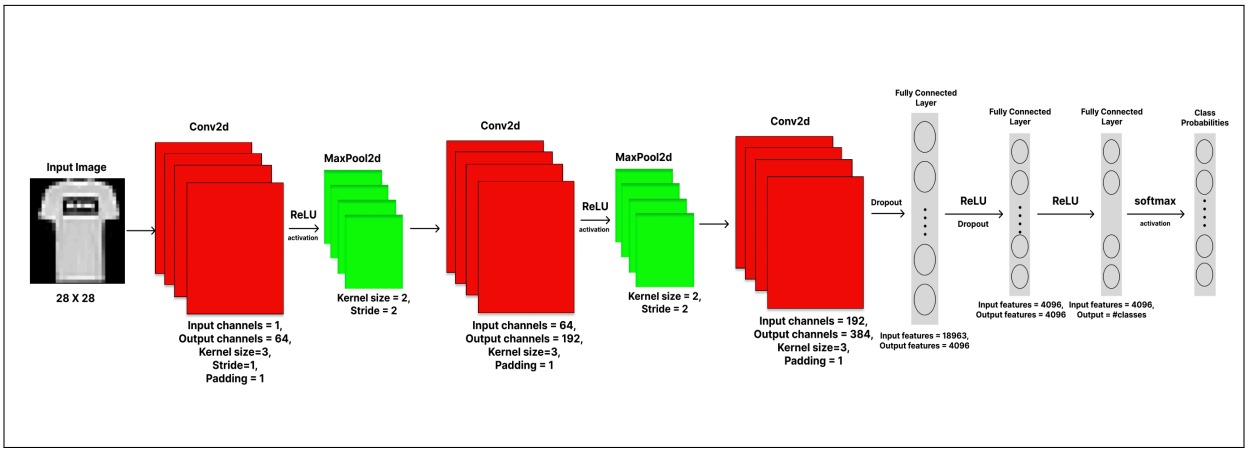


Figure 5: The modified AlexNet architecture used in the implementation of the experiments. Three convolution layers, first two followed by ReLU activation and max pool layers were used. Finally, three fully connected layers were used to classify the images and get output probabilities from the model.

4.3.3 Ensemble of AlexNet and LeNet

The ensemble method combines the strengths of both LeNet and AlexNet to achieve better generalization. By averaging the output probabilities of the final layer from both models, the ensemble captures a more robust representation of the data. The ensembling can be mathematically represented by Equation 17 as:

$$p_{ensemble}(y|x) = \frac{1}{2}\left(p_{LeNet}(y|x) + p_{AlexNet}(y|x)\right), \tag{17}$$

where $p_{ensemble}(y|x)$ is the probability of the label y given input x for the ensemble, and $p_{AlexNet}(y|x)$ and $p_{LeNet}(y|x)$ are the probabilities given by the individual AlexNet and LeNet models, respectively.

This averaging process smooths out the predictions, making the ensemble less likely to overfit to noise in the dataset and often resulting in improved classification accuracy on test data. The use of these models and their ensemble as teacher models in our experiments aims to provide comprehensive learning signals for the student QNN, leveraging the classical CNNs' ability to extract hierarchical features from image data.

4.4 Parameterized Quantum Circuits

In our study, we employed two Parameterized Quantum Circuits (PQCs): one with 3 qubits and another with 4 qubits. PQCs are central to quantum machine learning and variational algorithms, where the parameters (often denoted as θ) of quantum gates are tuned to minimize a cost function. The PQCs that are constructed and used in our experiments are shown in Figure 6. The 3-qubit PQC is constructed using Hadamard gates (H gates) and rotation gates around the y-axis (Ry gates), which are parameterized by angles θ. These angles are randomly initialized and optimized during the learning process. The circuit provides a compact yet expressive model that can capture the complexity needed for smaller-scale quantum tasks. Similarly, the 4-qubit PQC consists of a sequence of Ry gates with randomly initialized parameters θ. With an additional qubit, the 4-qubit circuit can represent a larger state space, allowing it to capture more complex patterns and correlations in the data.

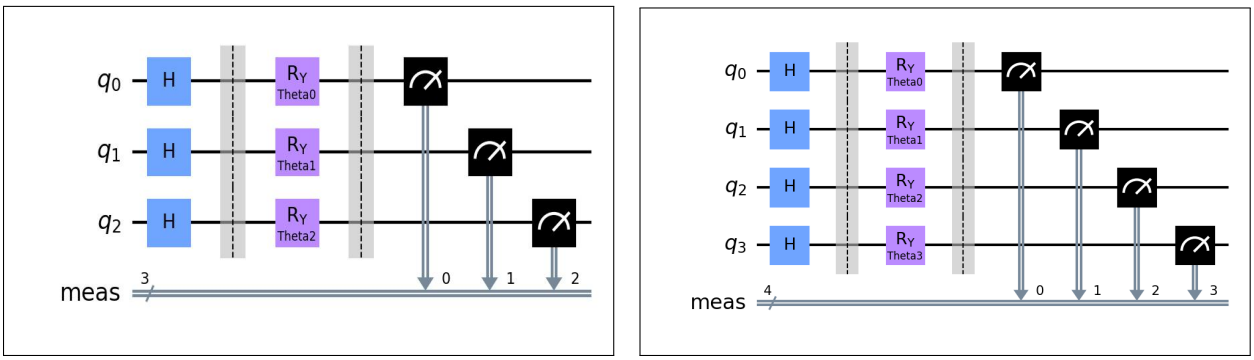


Figure 6: Parameterized Quantum Circuits used in the experiments. (a) A simple 3-qubit PQC with H and Ry gates parameterized by θ. (b) A simple 4-qubit PQC with H and Ry gates parameterized by θ.

The use of these two circuits is vital with respect to the flexibility and scalability of the methodology. A 3-qubit model demonstrates how our method performs with a minimal number of qubits, which is important for near-term quantum devices with limited qubit resources. The 4-qubit model, on the other hand, tests the capability of our method to scale to a slightly larger system, which is crucial for assessing performance in more demanding tasks. Together, they demonstrate generalizability and broader application potential across different quantum system sizes, which is essential for practical quantum machine learning implementations.

4.5 Performance Metrics

To evaluate the performance of our quantum and classical models, we rely on accuracy as the primary metric. Accuracy is a suitable measure in the context of balanced datasets, such as the ones we have crafted through our filtration process. The accuracy of a model is calculated as the ratio of correctly predicted instances to the total number of instances, as shown in Equation 18:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}. \tag{18}$$

Since our filtered datasets are balanced with an equal number of samples from each class, accuracy serves as a reliable metric. In datasets where classes are imbalanced, metrics such as precision, recall, or the F1 score might be more appropriate. However, for balanced datasets, accuracy straightforwardly reflects the model's capability to classify new data points correctly. Accuracy is not only a clear and interpretable metric but also practical from a computational standpoint. It allows us to clearly demonstrate the effectiveness of our models without incurring additional computational costs or complexity. Accuracy provides a direct measure of a model's classification performance and is a widely used metric in machine learning research for balanced datasets, making it an appropriate choice for this study.

4.6 Implementation Details

In our experimental setup, we employed PyTorch [58] as the framework for implementing neural network models, with the Qiskit QASM Simulator [59] used for conducting quantum simulations. Hyperparameter optimization was used, enabling the selection of optimal values for the optimizer, learning rate, loss function, temperature for KL-Divergence, and lambda (λ) (Equation 7). The Adam optimizer [51] was chosen for its effectiveness in handling noisy problems and sparse gradients. We set the learning rate to 0.001 and limited the training to 10 epochs to prevent overfitting while allowing sufficient model training. The loss function employed was Cross-Entropy Loss [46] for classification, complemented by KL-Divergence for the knowledge distillation process, with a temperature setting of 3 to soften the output distributions. A λ value of 0.9 was utilized to weigh the distillation loss in the overall loss function appropriately. For the quantum experiments, we performed 5000 shots for each measurement to approximate the expectation values with high accuracy and interpret them as probability distributions. These hyperparameters were meticulously chosen based on extensive hyperparameter optimization to ensure the best performing models.

4.7 Results

Table 1: Performance of Quantum and Classical Architectures on MNIST dataset variants. The table is divided into Student, Teacher, and Distillation categories to exhibit the direct training, teacher, and knowledge distillation results respectively.

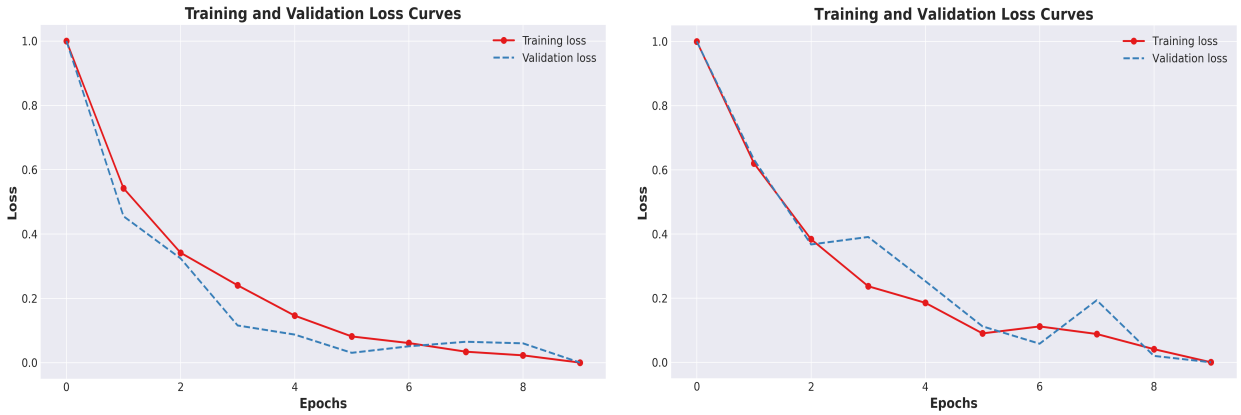
Category	Architecture	Dataset	Accuracy (%)	+ Δ for Student
Student	Quantum, 3 qubit	MNIST 0-5	96.56	-
	Quantum, 3 qubit	MNIST 0-7	93.75	-
	Quantum, 4 qubit	MNIST 0-7	93.83	-
Teacher	Classical, AlexNet	MNIST 0-5	99.72	-
	Classical, AlexNet	MNIST 0-7	99.51	-
	Classical, LeNet	MNIST 0-5	99.50	-
	Classical, LeNet	MNIST 0-7	99.16	-
	Classical, Ensemble	MNIST 0-5	99.73	-
	Classical, Ensemble	MNIST 0-7	99.58	-
Distillation	Teacher: AlexNet, Student: Quantum, 3 qubit	MNIST 0-5	97.44	0.88%
	Teacher: AlexNet, Student: Quantum, 3 qubit	MNIST 0-7	94.92	1.17%
	Teacher: AlexNet, Student: Quantum, 4 qubit	MNIST 0-7	94.17	0.34%
	Teacher: LeNet, Student: Quantum, 3 qubit	**MNIST 0-5**	**97.89**	**1.33%**
	Teacher: LeNet, Student: Quantum, 3 qubit	MNIST 0-7	93.75	0.00%
	Teacher: LeNet, Student: Quantum, 4 qubit	**MNIST 0-7**	**94.33**	**0.50%**
	Teacher: Ensemble, Student: Quantum, 3 qubit	MNIST 0-5	97.33	0.77%
	Teacher: Ensemble, Student: Quantum, 3 qubit	**MNIST 0-7**	**95.42**	**1.67%**
	Teacher: Ensemble, Student: Quantum, 4 qubit	MNIST 0-7	94.00	0.17%

The results of the conducted experiments are presented in the two comprehensive tables Table 1 and Table 2, which detail the performance of various architectures across the filtered MNIST and FashionMNIST datasets. The results are categorized into three distinct groups: Student, Teacher, and Distillation, to demonstrate the effectiveness of knowledge transfer in quantum neural network training. The 'Student' category represents the quantum models directly trained on the datasets. The 'Teacher' category includes the performance of classical convolutional neural networks, serving as the teacher models to guide the students during distillation. Lastly, the 'Distillation' category shows the outcomes of the student quantum models when knowledge is distilled from the classical teacher models, highlighting the enhancement in performance due to the distillation process. The "+ Δ for Student" represents the positive change in accuracy of the student quantum models upon training by knowledge distillation. The Architecture column shows the kind of model used, and in case of knowledge distillation, the teacher and student models used in each experiment. In case of the Student and Teacher categories, the positive increase for student is not relevant, hence kept blank using '-'. As depicted in Figure 7, the loss curves for the distillation training of the two student models for 3 and 4 qubits for the FashionMNIST subsets are presented. Each subfigure illustrates the decreasing trend of training and validation losses over epochs, indicating effective learning and generalization of the models. The consistent decline in loss values across all models signifies the models' ability to learn and adapt from the distilled knowledge, underlining the potential of quantum neural networks in assimilating complex patterns from classical pre-trained networks. The convergence

Table 2: Performance of Quantum and Classical Architectures on FashionMNIST dataset variants. Similar to MNIST, the results are divided into the three categories to provide insights into the quantum models' capabilities with and without the influence of knowledge distillation.

Category	Architecture	Dataset	Accuracy (%)	+ Δ for Student
Student	Quantum, 3 qubit	FashionMNIST 0-5	65.00	-
	Quantum, 3 qubit	FashionMNIST 0-7	68.00	-
	Quantum, 4 qubit	FashionMNIST 0-7	74.25	-
Teacher	Classical, AlexNet	FashionMNIST 0-5	94.90	-
	Classical, AlexNet	FashionMNIST 0-7	91.64	-
	Classical, LeNet	FashionMNIST 0-5	91.57	-
	Classical, LeNet	FashionMNIST 0-7	85.97	-
	Classical, Ensemble	FashionMNIST 0-5	95.67	-
	Classical, Ensemble	FashionMNIST 0-7	91.79	-
Distillation	**Teacher: AlexNet, Student: Quantum, 3 qubit**	**FashionMNIST 0-5**	**83.44**	**18.44%**
	Teacher: AlexNet, Student: Quantum, 3 qubit	FashionMNIST 0-7	73.42	5.42%
	Teacher: AlexNet, Student: Quantum, 4 qubit	FashionMNIST 0-7	75.42	1.17%
	Teacher: LeNet, Student: Quantum, 3 qubit	FashionMNIST 0-5	77.00	12.00%
	Teacher: LeNet, Student: Quantum, 3 qubit	FashionMNIST 0-7	68.67	0.67%
	Teacher: LeNet, Student: Quantum, 4 qubit	**FashionMNIST 0-7**	**75.83**	**1.58%**
	Teacher: Ensemble, Student: Quantum, 3 qubit	FashionMNIST 0-5	67.89	2.89%
	Teacher: Ensemble, Student: Quantum, 3 qubit	**FashionMNIST 0-7**	**74.42**	**6.42%**
	Teacher: Ensemble, Student: Quantum, 4 qubit	FashionMNIST 0-7	72.50	0.00%

of these curves demonstrates the learning of the models, indicative of the successful transfer of knowledge from the teacher models to their quantum counterparts.

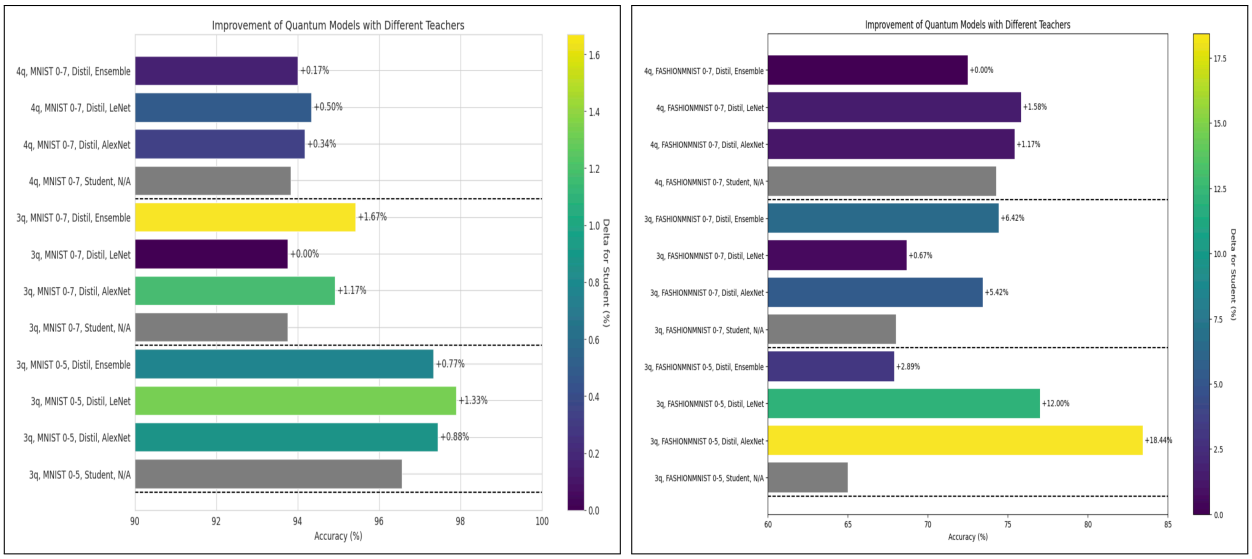


(a) (b)

Figure 7: Loss curves for distillation training of the two best student models for the FashionMNIST dataset, showcasing the efficiency of knowledge transfer. (a) Represents the learning curve of the best performing 3-qubit model with FashionMNIST 0-5 and AlexNet as the teacher model. (b) Represents the best performing 4-qubit model with FashionMNIST 0-7 and LeNet as the teacher model.

5 Findings and Discussion

Based on the experimental results of the previous section, in this section, we present our findings and discuss upon them. Based on Table 1, Table 2 and Figure 8, we can observe the followings: **(1)** When used as a teacher model in knowledge distillation, the LeNet and ensemble models exhibit superior performance compared to the simplified AlexNet architecture. This is likely because LeNet was explicitly designed for grayscale images like those in the



(a) (b)

Figure 8: Bar plot visually showing the improvement of the student quantum models compared to each distillation experiment performed on the (a) MNIST dataset subsets and (b) FashionMNIST dataset subsets. Each plot is divided into three groups for the three student quantum models. The grey colored bar in each group represents the baseline performance of the student models for the subset mentioned. Subsequent bars represent the improvement for each of the teacher models AlexNet, LeNet and Ensemble model.

MNIST and FashionMNIST datasets, allowing it to capture features more effectively than the more complex AlexNet, which required simplification for our experiments. **(2)** We observe that the average improvement in performance on the FashionMNIST dataset is greater than that on the MNIST dataset. This could be due to FashionMNIST's inherently lower baseline accuracies, which leaves more room for improvement. This disparity also highlights the third observation: **(3)** FashionMNIST, with its more varied and complex images, poses a greater classification challenge than MNIST. Yet, our knowledge distillation (KD) approach manages to improve the performance of quantum models on FashionMNIST significantly, indicating the robustness and applicability of KD in enhancing models trained on complex datasets. **(4)** It can also be observed that the 4-qubit student models do not significantly outperform the 3-qubit models. This suggests that increasing the complexity of the quantum model slightly does not necessarily lead to better learning. This is a testament to our main research objective: reducing the need for quantum resources while improving and maintaining robust performance. **(5)** The ensemble model performs best for the MNIST 0-7 and FashionMNIST 0-7 datasets, which are the most extensive datasets in our experiments. This model's superior performance is attributed to its ability to integrate diverse features and learning aspects from multiple classical models, thereby enhancing its predictive accuracy. Interestingly, this advantage is more highlighted in larger datasets. In contrast, the performance gains in smaller datasets are less remarkable, suggesting that ensemble models excel when there is a wealth of variation to learn from, while in more limited datasets, the opportunity to leverage their full potential is not as readily available. **(6)** Lastly, we note that the performance improvements through distillation are consistent across different dataset filtrations. This consistency underscores the potential for KD to be used in various quantum machine learning applications, especially as quantum hardware continues to evolve and allows for training on larger and more diverse datasets.

The findings signify the potential for classical machine learning architectures to substantially enhance the performance of quantum models through knowledge distillation. The compatibility of the teacher model with the data is crucial, as evidenced by the superior performance of LeNet and ensemble models tailored for grayscale image datasets. The advancements in knowledge distillation demonstrated in this study are particularly promising for complex datasets, suggesting that as quantum computational resources grow, the scope of quantum machine learning applications will broaden significantly. The use of established classical models like ResNet [60], VGG [61], and InceptionNet [62] as teachers in quantum environments might open up new possibilities for cross-domain applications, reducing the reliance on the development of large quantum models and leveraging the extensive resources of classical pre-trained models. This research paves the way for future endeavors where quantum models are integrated into everyday tasks, marking a significant step forward in the evolution of quantum computing applications.

6 Conclusion and Future Work

In this paper, we presented a new method for transferring knowledge from classical neural networks to quantum neural networks via knowledge distillation, showcasing a promising avenue for leveraging classical machine learning architectures within quantum computing paradigms. By tweaking and leveraging classical models like LeNet, AlexNet and their ensemble to act as teachers, we demonstrated significant improvements in quantum model performances on both MNIST and FashionMNIST datasets, with higher enhancements observed in the more complex FashionMNIST dataset. The results showed an average improvement of 0.80% on the MNIST dataset and 5.40% on the FashionMNIST dataset.

Our results indicate the potential for classical-to-quantum knowledge distillation in a variety of applications. In natural language processing tasks, where data representation and model interpretability are paramount, the method could be utilized in developing more efficient quantum models. The healthcare sector could also benefit immensely, with applications ranging from disease prediction to clinical decision-making, where quantum-enhanced models could lead to faster and more accurate diagnoses. Furthermore, the principles of this research hold great promise for advancing object identification, tracking, and detection tasks, crucial in areas such as autonomous driving and surveillance. As quantum computing continues to evolve, the techniques outlined in this study are of paramount importance in bridging the gap between classical machine learning successes and quantum computing's potential, fostering advancements across an array of domains and setting a precedent for future quantum machine learning research.

References

[1] Rafał Potempa and Sebastian Porebski. Comparing concepts of quantum and classical neural network models for image classification task. In *Progress in Image Processing, Pattern Recognition and Communication Systems: Proceedings of the Conference (CORES, IP&C, ACS)-June 28-30 2021 12*, pages 61–71. Springer, 2022.

[2] Mahabubul Alam and Swaroop Ghosh. Qnet: A scalable and noise-resilient quantum neural network architecture for noisy intermediate-scale quantum computers. *Frontiers in Physics*, 9:702, 2022.

[3] Raiyan Rahman, Md Shawmoon Azad, Mohammed Rakibul Hasan, Syed Emad Uddin Shubha, and MRC Mahdy. Enhancing the security of image transmission in quantum era: A chaos-assisted qkd approach using entanglement. *arXiv preprint arXiv:2311.18471*, 2023.

[4] Mahabubul Alam, Satwik Kundu, and Swaroop Ghosh. Knowledge distillation in quantum neural network using approximate synthesis. In *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, pages 639–644, 2023.

[5] Sungho Shin, Yoonho Boo, and Wonyong Sung. Knowledge distillation for optimization of quantized deep neural networks. In *2020 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 1–6. IEEE, 2020.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[8] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[9] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[10] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[11] Lars S Madsen, Fabian Laudenbach, Mohsen Falamarzi Askarani, Fabien Rortais, Trevor Vincent, Jacob FF Bulmer, Filippo M Miatto, Leonhard Neuhaus, Lukas G Helt, Matthew J Collins, et al. Quantum computational advantage with a programmable photonic processor. *Nature*, 606(7912):75–81, 2022.

[12] Peter W Shor. Scheme for reducing decoherence in quantum computer memory. *Physical review A*, 52(4):R2493, 1995.

[13] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, 2022.

[14] AMJ Zwerver, T Krähenmann, TF Watson, Lester Lampert, Hubert C George, Ravi Pillarisetty, SA Bojarski, Payam Amin, SV Amitonov, JM Boter, et al. Qubits made by advanced semiconductor manufacturing. *Nature Electronics*, 5(3):184–190, 2022.

[15] Hajime Okamoto, Adrien Gourgout, Chia-Yuan Chang, Koji Onomitsu, Imran Mahboob, Edward Yi Chang, and Hiroshi Yamaguchi. Coherent phonon manipulation in coupled mechanical resonators. *Nature Physics*, 9(8):480–484, 2013.

[16] Mohammad H Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytskyy, and Roger Melko. Quantum boltzmann machine. *Physical Review X*, 8(2):021050, 2018.

[17] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195–202, 2017.

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

[19] 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.

[20] Valeria Saggio, Beate E Asenbeck, Arne Hamann, Teodor Strömberg, Peter Schiansky, Vedran Dunjko, Nicolai Friis, Nicholas C Harris, Michael Hochberg, Dirk Englund, et al. Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849):229–233, 2021.

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

[22] Adam Bouland, Wim van Dam, Hamed Joorati, Iordanis Kerenidis, and Anupam Prakash. Prospects and challenges of quantum finance. *arXiv preprint arXiv:2011.06492*, 2020.

[23] Ryuji Takagi, Suguru Endo, Shintaro Minagawa, and Mile Gu. Fundamental limits of quantum error mitigation. *npj Quantum Information*, 8(1):114, 2022.

[24] Dayue Qin, Yanzhu Chen, and Ying Li. Error statistics and scalability of quantum error mitigation formulas. *npj Quantum Information*, 9(1):35, 2023.

[25] Phalgun Lolur, Mårten Skogh, Werner Dobrautz, Christopher Warren, Janka Biznárová, Amr Osman, Giovanna Tancredi, Goran Wendin, Jonas Bylander, and Martin Rahm. Reference-state error mitigation: A strategy for high accuracy quantum computation of chemistry. *Journal of Chemical Theory and Computation*, 19(3):783–789, 2023.

[26] Abhinav Kandala, Kristan Temme, Antonio D Córcoles, Antonio Mezzacapo, Jerry M Chow, and Jay M Gambetta. Error mitigation extends the computational reach of a noisy quantum processor. *Nature*, 567(7749):491–495, 2019.

[27] Armands Strikis, Dayue Qin, Yanzhu Chen, Simon C Benjamin, and Ying Li. Learning-based quantum error mitigation. *PRX Quantum*, 2(4):040330, 2021.

[28] Daniel Bultrini, Max Hunter Gordon, Piotr Czarnik, Andrew Arrasmith, M Cerezo, Patrick J Coles, and Lukasz Cincio. Unifying and benchmarking state-of-the-art quantum error mitigation techniques. *Quantum*, 7:1034, 2023.

[29] Angus Lowe, Max Hunter Gordon, Piotr Czarnik, Andrew Arrasmith, Patrick J Coles, and Lukasz Cincio. Unified approach to data-driven quantum error mitigation. *Physical Review Research*, 3(3):033098, 2021.

[30] Anita Weidinger, Glen Bigan Mbeng, and Wolfgang Lechner. Error mitigation for quantum approximate optimization. *arXiv preprint arXiv:2301.05042*, 2023.

[31] Marco Fellous-Asiani, Jing Hao Chai, Yvain Thonnart, Hui Khoon Ng, Robert S Whitney, and Alexia Auffèves. Optimizing resource efficiencies for scalable full-stack quantum computers. *PRX Quantum*, 4(4):040319, 2023.

[32] Margaret Martonosi and Martin Roetteler. Next steps in quantum computing: Computer science's role. *arXiv preprint arXiv:1903.10541*, 2019.

[33] Olawale Ayoade, Pablo Rivas, and Javier Orduz. Artificial intelligence computing at the quantum level. *Data*, 7(3):28, 2022.

[34] Annarita Giani and Zachary Eldredge. Quantum computing opportunities in renewable energy. *SN Computer Science*, 2(5):393, 2021.

[35] Remmy Zen, Long My, Ryan Tan, Frédéric Hébert, Mario Gattobigio, Christian Miniatura, Dario Poletti, and Stéphane Bressan. Transfer learning for scalability of neural-network quantum states. *Physical Review E*, 101(5):053301, 2020.

[36] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian Roitberg. Outsmarting quantum chemistry through transfer learning. 2018.

[37] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):2903, 2019.

[38] Andrea Mari, Thomas R Bromley, Josh Izaac, Maria Schuld, and Nathan Killoran. Transfer learning in hybrid classical-quantum neural networks. *Quantum*, 4:340, 2020.

[39] Muhammad Junaid Umer, Javeria Amin, Muhammad Sharif, Muhammad Almas Anjum, Faisal Azam, and Jamal Hussain Shah. An integrated framework for covid-19 classification based on classical and quantum transfer learning from a chest radiograph. *Concurrency and Computation: Practice and Experience*, 34(20):e6434, 2022.

[40] Mine Kaya and Shima Hajimirza. Using a novel transfer learning method for designing thin film solar cells with enhanced quantum efficiencies. *Scientific reports*, 9(1):5034, 2019.

[41] Javeria Amin, Muhammad Almas Anjum, Muhammad Sharif, Saima Jabeen, Seifedine Kadry, Pablo Moreno Ger, et al. A new model for brain tumor detection using ensemble transfer learning and quantum variational classifier. *Computational intelligence and neuroscience*, 2022, 2022.

[42] T Tamilvizhi, R Surendran, K Anbazhagan, and K Rajkumar. Quantum behaved particle swarm optimization-based deep transfer learning model for sugarcane leaf disease detection and classification. *Mathematical Problems in Engineering*, 2022, 2022.

[43] Dipendra Jha, Kamal Choudhary, Francesca Tavazza, Wei-keng Liao, Alok Choudhary, Carelyn Campbell, and Ankit Agrawal. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature communications*, 10(1):5316, 2019.

[44] Jun Qi and Javier Tejedor. Classical-to-quantum transfer learning for spoken command recognition based on quantum neural networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8627–8631. IEEE, 2022.

[45] Juhyeon Kim, Joonsuk Huh, and Daniel K Park. Classical-to-quantum convolutional neural network transfer learning. *Neurocomputing*, 555:126643, 2023.

[46] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[47] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[48] John Von Neumann. *Mathematical foundations of quantum mechanics: New edition*, volume 53. Princeton university press, 2018.

[49] Jin-Guo Liu and Lei Wang. Differentiable learning of quantum circuit born machines. *Physical Review A*, 98(6):062324, 2018.

[50] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[51] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[52] Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. In *International conference on machine learning*, pages 2545–2553. PMLR, 2017.

[53] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[54] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504, 2019.

[55] Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002*, 2018.

[56] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[57] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[59] Gadi Aleksandrowicz, Thomas Alexander, Panagiotis Barkoutsos, Luciano Bello, Yael Ben-Haim, David Bucher, F Jose Cabrera-Hernández, Jorge Carballo-Franquis, Adrian Chen, Chun-Fu Chen, et al. Qiskit: An open-source framework for quantum computing. *Accessed on: Mar*, 16, 2019.

[60] 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*, pages 770–778, 2016.

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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