

Quantum Kernels for Audio Deepfake Detection Using Spectrogram Patch Features

Lisan Al Amin*, Rakib Hossain*, Mahbubul Islam[†], Faisal Quader[‡], and Thanh Thi Nguyen^{§¶}

*Potomac Quantum, USA

[†]United International University, Dhaka, Bangladesh

[‡]University of Maryland, College Park, MD, USA

[§]Monash University, Melbourne, Victoria, Australia

[¶]University of the Sunshine Coast, Queensland, Australia

E-mails: lisanalamin@gmail.com, rakib.sat18@gmail.com, mahbub120.eee@gmail.com, fquader1@umbc.edu, nguyens5@usc.edu.au

Abstract—Quantum machine learning has emerged as a promising tool for pattern recognition, yet many audio-focused approaches still treat spectrograms as generic images and do not explicitly exploit their time-frequency structure. We propose Q-Patch, a quantum feature map tailored to audio that encodes local time-frequency patches from mel-spectrograms into quantum states using shallow, hardware-efficient circuits with adjacency-aware entanglement. Each selected patch is summarized by a compact four-dimensional acoustic descriptor and mapped to a four-qubit circuit with depth at most three, enabling practical quantum kernel construction under near-term constraints. We evaluate Q-Patch on an audio spoofing detection task using a controlled, balanced protocol and compare it with size-matched classical baselines. Q-Patch improves discrimination between bona fide and spoofed samples, achieving an area under the receiver operating characteristic curve (AUROC) of 0.87, compared with 0.82 for a radial basis function support vector machine (RBF-SVM) trained on the same patch-level features. Kernel-space analysis further reveals a clear class structure, with cross-class similarity around 0.615 and within-class self-similarity of 1.00. Overall, Q-Patch provides a practical framework for incorporating time-frequency-aware representations into quantum kernel learning for audio authenticity assessment in low-resource settings.

Index Terms—Quantum machine learning, spectrogram analysis, audio deepfake detection, anti-spoofing, robustness, few-shot learning

I. INTRODUCTION

Recent advances in generative speech have made modern text-to-speech (TTS) and voice conversion (VC) systems capable of producing audio that is often difficult to distinguish from genuine human speech. While these technologies support accessibility and content creation, they also introduce serious security risks, including impersonation, fraud, and misinformation. As a result, reliable audio spoofing detection has become an important research problem [1].

Benchmarks such as ASVspoof 2019 [2] and ADD 2022 [3] have accelerated progress, but many detectors still struggle under unseen attacks and real-world distortions, including channel effects, background noise, compression, and replay artifacts [4], [5]. In addition, many existing approaches process spectrograms as generic images, which can overlook important time-frequency structure in speech.

Quantum machine learning offers an alternative framework for representation learning and similarity estimation through quantum state overlaps [6]. In particular, quantum kernel methods are attractive in limited-data settings, yet their use in speech security remains limited, and most existing pipelines are not designed around the structure of time-frequency audio representations [7].

Motivated by these gaps, we propose *Q-Patch*, a quantum feature-mapping framework for audio spoofing detection that encodes informative local time-frequency patches into shallow quantum circuits. The main contributions of this work are as follows:

- We introduce a patch-based audio representation that summarizes local time-frequency regions using compact and interpretable acoustic descriptors.
- We design a shallow quantum feature map with adjacency-aware entanglement, limiting circuit depth to three layers and qubit usage to 4–8 qubits for compatibility with Noisy Intermediate-Scale Quantum (NISQ) constraints.
- We evaluate the proposed framework in a controlled low-resource setting against size-matched classical baselines and analyze the induced kernel space using both classification metrics and similarity structure.

We validate Q-Patch on controlled audio data derived from LJ Speech [8]. The results suggest that time-frequency-aware quantum feature maps can provide a useful inductive bias for audio authenticity discrimination while maintaining a shallow circuit design suitable for near-term quantum settings.

The rest of the paper is organized as follows. Section II reviews related work, Section III presents the proposed methodology and experimental protocol, whilst Section IV reports the results and discussion, and Section V concludes the paper.

II. RELATED WORK

This section reviews prior work in audio deepfake detection and quantum kernel learning, and then highlights the gap addressed by *Q-Patch*.

A. Audio Deepfake Detection and Anti-spoofing

The increasing realism of TTS and VC systems has driven substantial progress in audio deepfake detection. Surveys by Khanjani *et al.* [1], [9] and Pham *et al.* [7] trace the field from handcrafted features to deep learning, supported by benchmarks such as ASVspoof 2019 [2] and ADD 2022 [3]. These benchmarks provide standardized protocols and include degraded and channel-distorted conditions that better reflect deployment scenarios.

Current state-of-the-art systems are dominated by deep neural architectures. RawNet2 [10] and DeepLASD [6], for example, learn discriminative representations directly from waveforms or spectrogram-like inputs. Related evidence also suggests that compact models can remain effective under strong resource constraints; Al Amin *et al.* [11] show that sparse subnetworks can retain strong detection capability. Related multimodal studies have also shown that self-supervised vision and speech encoders can provide effective compact representations in applied human-centered settings [12], reinforcing the broader case for structured feature design beyond large end-to-end architectures. Despite these advances, robustness remains a key challenge, with performance often degrading under unseen attacks, distribution shifts, and realistic distortions [4], [5]. This motivates approaches that improve generalization and better exploit time–frequency structure.

B. Quantum Machine Learning and Kernel Methods

Quantum machine learning investigates whether quantum feature representations and quantum-native similarity measures can improve learning performance. In the context of kernel methods, Schnabel and Roth [13] show that feature-map design plays a central role in kernel expressivity. Egginger *et al.* [14] and Innan *et al.* [15] further emphasize the importance of circuit depth, encoding strategy, and trainability in quantum kernel quality.

Applications of quantum kernels are beginning to emerge across domains. Beaulieu *et al.* [16] apply them to manufacturing defect detection, while Tran *et al.* [17] explore their use in biomedical speech analysis. Broader surveys [18], [19] also stress the importance of shallow, hardware-aware designs in the Noisy Intermediate-Scale Quantum era. However, quantum-kernel methods remain underexplored for audio deepfake detection, especially when the representation is designed explicitly around spectrogram time–frequency structure.

C. Gaps and Motivation

The literature points to three main gaps. First, robustness to unseen attacks and real-world distortions remains limited [4], [5]. Second, low-resource detection settings are still less studied, despite their practical relevance when new spoofing methods appear. Third, the use of quantum kernels for audio deepfake detection remains limited, and existing quantum machine learning approaches rarely incorporate inductive biases tailored to spectrogram geometry [20]. These gaps motivate *Q-Patch*, a time–frequency-aware quantum kernel framework

designed for structured spectrogram patches, limited-data discrimination, and compatibility with near-term quantum hardware.

III. METHODOLOGY

This section describes the proposed *Q-Patch* framework, including data preparation, patch-level feature construction, quantum embedding, training procedure, baseline design, and implementation details.

A. Overview

We introduce *Q-Patch*, a quantum feature-mapping framework designed to capture local time–frequency structure in speech spectrograms using shallow entangling circuits. The method follows three stages: (i) computing an utterance-level log-mel representation, (ii) partitioning the time–frequency plane into small patches and summarizing each patch with a compact vector of interpretable statistics, and (iii) embedding the selected patch summaries into quantum states to form a fidelity-based kernel for a Quantum Support Vector Machine (QSVM) [14], [16].

To isolate the contribution of the proposed quantum feature map, we compare *Q-Patch* with matched classical baselines, including a radial basis function support vector machine (RBF-SVM) trained on the same patch summaries and a compact convolutional neural network (CNN) with $\leq 100k$ parameters trained directly on spectrograms. Algorithm 1, presented later in this section, specifies preprocessing, patching, and feature construction, whereas Algorithm 2 details quantum-kernel computation and QSVM training/inference.

Fig. 1 summarizes the end-to-end framework, from data preparation and patch summarization to quantum embedding, kernel learning, and evaluation.

B. Dataset and Experimental Setup

Due to the computational cost of quantum-kernel simulation, we construct a balanced 100-sample subset from LJ Speech [8], consisting of 50 bona fide utterances and 50 spoofed utterances. The spoofed samples are generated from the original recordings using additive Gaussian noise and spectral distortions as a controlled proxy for authenticity manipulation. We then partition this 100-sample set into 80 training and 20 development (dev) samples with preserved class balance and no overlap between splits. This compact protocol supports repeated kernel simulation and controlled comparison with classical baselines, but it should be interpreted as a feasibility study rather than evidence of broad real-world generalization.

C. Front-End Processing

Each waveform $x[n]$ is resampled to 16 kHz and converted into a log-mel spectrogram $M \in \mathbb{R}^{T \times F}$ with $F = 64$ mel bins. We compute the short-time Fourier transform (STFT) using a 25 ms Hann window and a 10 ms hop, with a 1024-point fast Fourier transform (FFT):

$$X(k, \tau) = \sum_n x[n] w[n - \tau] e^{-j2\pi kn/N}, \quad (1)$$

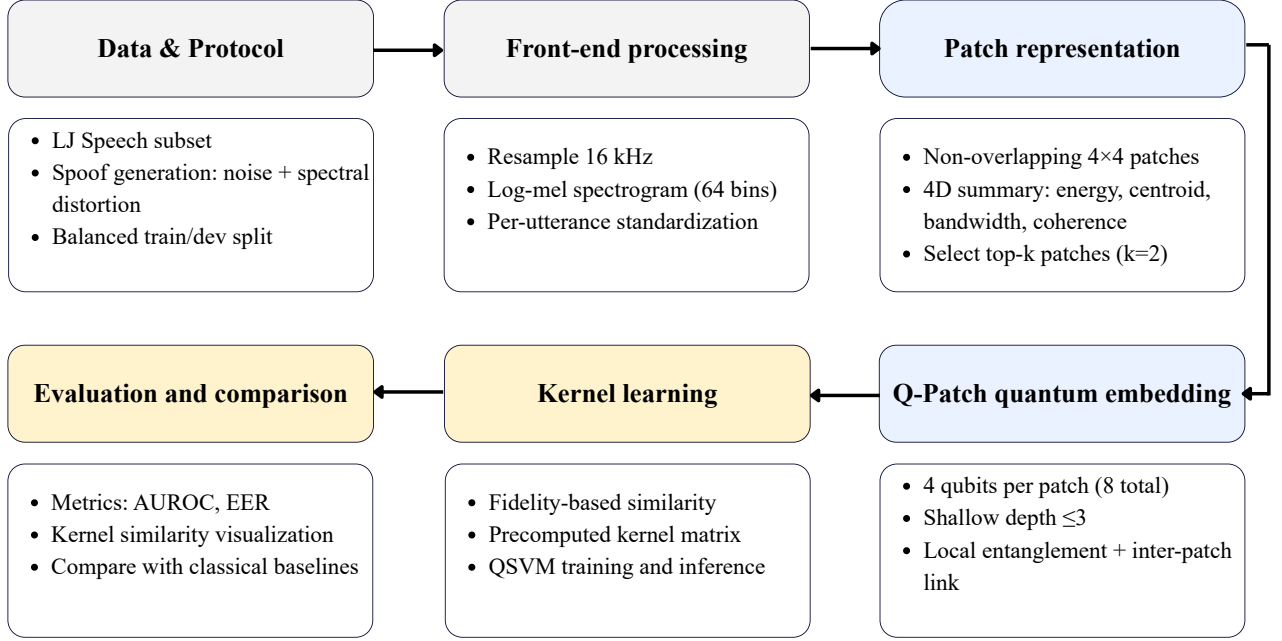


Fig. 1: Overview of the Q-Patch pipeline from data construction and time–frequency patch summarization to quantum embedding, kernel learning with QSVM, and evaluation against matched classical baselines.

where $w[\cdot]$ is the Hann window and $N = 1024$. Mel filterbank energies are then obtained as

$$E(f, \tau) = \sum_k |X(k, \tau)|^2 H_f(k), \quad (2)$$

where $H_f(\cdot)$ denotes the f -th mel filter. Log compression yields

$$M(\tau, f) = \log(E(f, \tau) + \epsilon), \quad (3)$$

with $\epsilon > 0$ for numerical stability. To reduce utterance-level scale variation, we apply per-utterance standardization:

$$\tilde{M}(\tau, f) = \frac{M(\tau, f) - \mu_M}{\sigma_M + \epsilon}, \quad (4)$$

where μ_M and σ_M are computed over all (τ, f) locations in the utterance. Fig. 2 shows representative bona fide and spoofed spectrograms after preprocessing.

D. Patch Partitioning and Summary Vectors

We partition the standardized spectrogram $\tilde{M} \in \mathbb{R}^{T \times F}$ into non-overlapping 4×4 time–frequency patches. Let $\mathcal{P} = \{P_i\}_{i=1}^{N_p}$ denote the patch set, with

$$P_i = \tilde{M}[\tau_i : \tau_i + 3, f_i : f_i + 3] \in \mathbb{R}^{4 \times 4}. \quad (5)$$

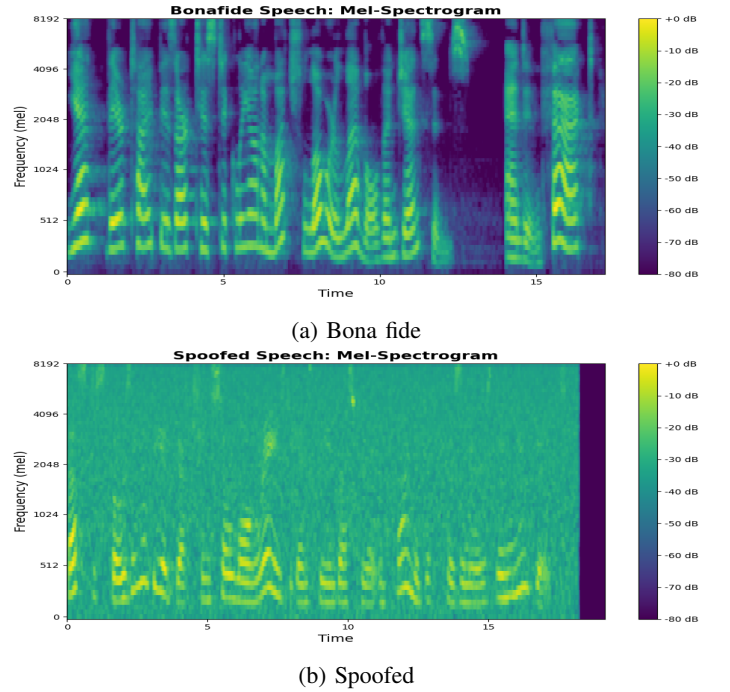


Fig. 2: Example spectrograms extracted from the LJ Speech dataset: (a) bona fide sample and (b) spoofed sample. Distinct local spectral patterns motivate patch-level modeling for classification.

Each patch P is mapped to a four-dimensional summary vector $\mathbf{s}(P) = [s_1, s_2, s_3, s_4]^\top$ that captures local intensity distribution and short-term temporal consistency:

$$s_1(P) = \frac{1}{|T||F|} \sum_{\tau \in T} \sum_{f \in F} P_{\tau,f} \quad (\text{Mean Patch Activation}), \quad (6)$$

$$\bar{m}_f(P) = \frac{1}{|T|} \sum_{\tau \in T} P_{\tau,f}, \quad (7)$$

$$w_f(P) = \frac{(|\bar{m}_f(P)| + \epsilon)}{\sum_{f' \in F} (|\bar{m}_{f'}(P)| + \epsilon)}, \quad (8)$$

$$s_2(P) = \sum_{f \in F} f w_f(P) \quad (\text{Spectral Centroid}), \quad (9)$$

$$s_3(P) = \sqrt{\sum_{f \in F} (f - s_2(P))^2 w_f(P)} \quad (\text{Spectral Bandwidth}), \quad (10)$$

$$s_4(P) = \frac{1}{|T| - 1} \sum_{\tau} \frac{\langle P_{\tau,:}, P_{\tau+1,:} \rangle}{\|P_{\tau,:}\|_2 \|P_{\tau+1,:}\|_2 + \epsilon} \quad (11)$$

(Inter-frame Coherence).

Here T indexes the four patch frames and F indexes the four patch frequency bins. We use $|\bar{m}_f(P)|$ in (8) to ensure nonnegative weights under per-utterance standardization.

a) Top- k patch selection.: To focus quantum resources on the most informative regions, we rank patches using a simple energy proxy and retain the top two. In our implementation, the ranking score is the mean activation $s_1(P_i)$:

$$\mathcal{I}_2 = \text{Top2} \left(\{s_1(P_i)\}_{i=1}^{N_p} \right). \quad (12)$$

We use $s_1(P_i)$ as a deterministic salience proxy that favors locally prominent time-frequency regions while keeping the selection rule training-free and computationally inexpensive under a strict qubit budget. We do not claim that mean activation is universally optimal; it may underweight quieter but discriminative regions, and sensitivity to alternative criteria, patch sizes, or small perturbations in spoof generation remains an important direction for future work. Under fixed preprocessing, however, the top- k selection is deterministic for each utterance.

The resulting feature vector is the concatenation of two patch summaries:

$$\mathbf{x} = [\mathbf{s}(P_{i_1}); \mathbf{s}(P_{i_2})] \in \mathbb{R}^8, \quad \{i_1, i_2\} = \mathcal{I}_2, \quad (13)$$

which maps naturally to an eight-qubit system, with four qubits assigned to each selected patch. Algorithm 1 summarizes the full feature-extraction and patch-selection procedure.

E. Quantum Feature Map (Q-Patch)

Given a patch summary vector $\mathbf{s} = [s_1, s_2, s_3, s_4]^\top$, Q-Patch embeds $\mathbf{s}$ into a four-qubit quantum state using parameterized single-qubit rotations followed by a lightweight entangling chain. Let $U_\phi(\mathbf{s})$ denote the embedding unitary:

$$U_{\text{rot}}(\mathbf{s}) = R_X(q_0, s_1) R_Y(q_1, s_2) R_Z(q_2, s_3) R_Y(q_3, s_4), \quad (14)$$

$$U_{\text{ent}} = CZ(q_0, q_1) CZ(q_1, q_2) CZ(q_2, q_3), \quad (15)$$

$$U_\phi(\mathbf{s}) = U_{\text{ent}} U_{\text{rot}}(\mathbf{s}), \quad (16)$$

$$|\phi(\mathbf{s})\rangle = U_\phi(\mathbf{s}) |0\rangle^{\otimes 4}. \quad (17)$$

Algorithm 1 Q-Patch feature extraction and top- k patch selection

Require: Audio waveform $x[n]$, sampling rate 16 kHz, mel bins $F = 64$, FFT size $N = 1024$, window 25 ms, hop 10 ms, patch size $(h, w) = (4, 4)$, top- $k = 2$

Ensure: Concatenated feature vector $\mathbf{x} \in \mathbb{R}^{4k}$

- 1: Compute log-mel spectrogram $M \in \mathbb{R}^{T \times F}$ from $x[n]$
- 2: Standardize per utterance: $\tilde{M} \leftarrow (M - \mu_M)/(\sigma_M + \epsilon)$
- 3: Partition $\tilde{M}$ into non-overlapping patches $\mathcal{P} = \{P_i\}_{i=1}^{N_p}$, each $P_i \in \mathbb{R}^{h \times w}$
- 4: **for** $i = 1$ to N_p **do**
- 5: Compute $\mathbf{s}(P_i) = [s_1, s_2, s_3, s_4]^\top$ using Eqs. (6)–(11)
- 6: Assign patch score $e_i \leftarrow s_1(P_i)$
- 7: **end for**
- 8: Select indices $\mathcal{I}_k \leftarrow \text{TopK}(\{e_i\}_{i=1}^{N_p}, k)$
- 9: Form $\mathbf{x} \leftarrow [\mathbf{s}(P_{i_1}); \dots; \mathbf{s}(P_{i_k})]$ where $\mathcal{I}_k = \{i_1, \dots, i_k\}$
- 10: **return** $\mathbf{x}$

a) Two-patch encoding (8 qubits).: For the top two patches, we apply the same embedding to two disjoint four-qubit blocks in parallel. Let $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ denote the selected patch summaries. The joint embedding is

$$|\phi(\mathbf{s}^{(1)}, \mathbf{s}^{(2)})\rangle = (CZ(q_3, q_4)) \left(U_\phi(\mathbf{s}^{(1)}) \otimes U_\phi(\mathbf{s}^{(2)}) \right) |0\rangle^{\otimes 8}, \quad (18)$$

where $CZ(q_3, q_4)$ introduces a single inter-patch interaction consistent with spatial adjacency. To remain compatible with Noisy Intermediate-Scale Quantum (NISQ) constraints [17], circuit depth is limited to $d \leq 3$. Single-qubit rotations are applied in parallel, and entangling operations are restricted to local controlled- Z chains, as illustrated in Fig. 3.

In all reported experiments, the two four-dimensional patch summaries in the concatenated feature vector are used directly as rotation angles in the embedding circuit; no additional learned rescaling or variational parameters are introduced. This preserves a controlled comparison with the RBF-SVM, since both methods operate on the same compact patch descriptor.

b) Kernel definition.: We define the quantum kernel between two inputs $\mathbf{x}$ and $\mathbf{x}'$ through state fidelity:

$$\kappa(\mathbf{x}, \mathbf{x}') = |\langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle|^2, \quad (19)$$

where $|\phi(\mathbf{x})\rangle$ is the embedded state constructed from the selected patch summaries. If more than two patches are used, we process non-overlapping patch pairs and define the final similarity as the arithmetic mean of the corresponding per-pair fidelities. Because nonnegative sums and averages of valid kernels remain positive semi-definite, this construction preserves QSVM compatibility while allowing scalability under bounded depth.

F. QSVM Training and Inference

Given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with labels $y_i \in \{-1, +1\}$, QSVM training uses the precomputed Gram matrix $K \in \mathbb{R}^{N \times N}$:

$$K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j), \quad (20)$$

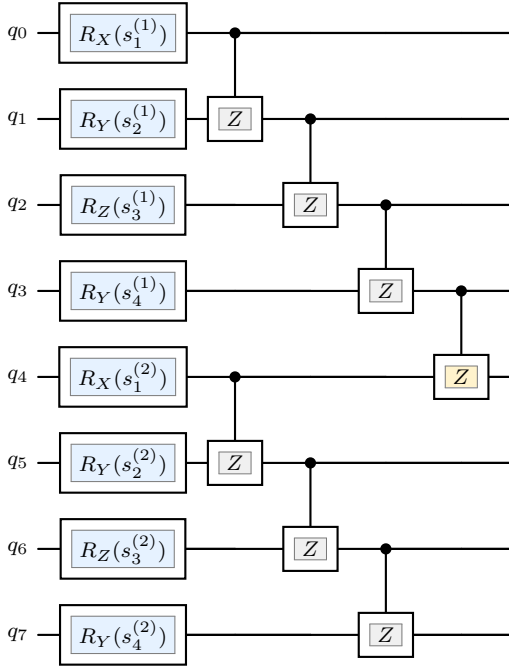


Fig. 3: Q-Patch feature map for two selected patches (8 qubits). Rotation encodings are shown in blue, controlled-Z targets in gray, and the inter-patch connection between q_3 and q_4 is highlighted in amber. The block can be repeated for depth $d \leq 3$.

where $\kappa(\cdot, \cdot)$ is defined in (19). At inference time, for a test example $\mathbf{x}^*$, we compute kernel values against the training set:

$$K_j^* = \kappa(\mathbf{x}^*, \mathbf{x}_j), \quad j = 1, \dots, N, \quad (21)$$

which corresponds to a standard kernel support vector machine decision function evaluated in the implicit feature space induced by the quantum embedding [14], [16]. Algorithm 2 details this procedure.

Unlike variational quantum classifiers, Q-Patch does not require iterative optimization of circuit parameters. Once the Gram matrix is computed, learning reduces to the standard convex support vector machine optimization with a precomputed kernel, so convergence is inherited from the underlying solver rather than from a separate non-convex quantum training loop.

G. Comparative Baselines

We evaluate Q-Patch against two baselines representing complementary classical modeling strategies.

a) *RBF-SVM*: A classical support vector machine with a radial basis function kernel is trained on the same patch-derived input $\mathbf{x} \in \mathbb{R}^8$, obtained by concatenating the top- k summary vectors. This baseline isolates the effect of the quantum feature map while holding the input representation fixed.

b) *Tiny CNN*: A compact convolutional neural network with $\leq 100k$ parameters is trained on spectrogram inputs. This baseline captures local time–frequency patterns without explicit patch summarization.

Algorithm 2 Quantum kernel construction and QSVM training/inference

Require: Training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $y_i \in \{-1, +1\}$, test set $\{\mathbf{x}_u^*\}_{u=1}^{N^*}$, embedding circuit $U_\phi(\cdot)$ with depth $d \leq 3$

Ensure: QSVM decision scores for test samples

```

1: function KERNEL( $\mathbf{x}, \mathbf{x}'$ )
2:   Prepare  $|\phi(\mathbf{x})\rangle \leftarrow U_\phi(\mathbf{x})|0\rangle^{\otimes n}$ 
3:   Prepare  $|\phi(\mathbf{x}')\rangle \leftarrow U_\phi(\mathbf{x}')|0\rangle^{\otimes n}$ 
4:   return  $\kappa(\mathbf{x}, \mathbf{x}') \leftarrow |\langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle|^2$ 
5: end function
6: Training kernel: initialize  $K \in \mathbb{R}^{N \times N}$ 
7: for  $i = 1$  to  $N$  do
8:   for  $j = 1$  to  $N$  do
9:      $K_{ij} \leftarrow \text{KERNEL}(\mathbf{x}_i, \mathbf{x}_j)$ 
10:  end for
11: end for
12: Train QSVM with precomputed kernel matrix  $K$  [14], [16]
13: Test kernel: initialize  $K^* \in \mathbb{R}^{N^* \times N}$ 
14: for  $u = 1$  to  $N^*$  do
15:   for  $j = 1$  to  $N$  do
16:      $K_{uj}^* \leftarrow \text{KERNEL}(\mathbf{x}_u^*, \mathbf{x}_j)$ 
17:   end for
18: end for
19: Compute QSVM decision scores using  $K^*$ 

```

c) *Evaluation metrics*: We report the area under the receiver operating characteristic curve (AUROC) and equal error rate (EER). AUROC summarizes ranking quality across all thresholds, while EER is standard in anti-spoofing and directly characterizes the false-accept/false-reject trade-off at a single operating point. EER is computed at the operating point τ^* where the false positive rate (FPR) equals the false negative rate (FNR):

$$\text{EER} = \text{FPR}(\tau^*) \quad \text{such that} \quad \text{FPR}(\tau^*) = \text{FNR}(\tau^*). \quad (22)$$

In addition, we analyze quantum-kernel similarity patterns to assess whether bona fide and spoofed samples form separable structure in the induced kernel space, with particular attention to feasibility under shallow depth constraints and limited-data conditions.

H. Implementation Notes

Circuit depth is constrained to $d \leq 3$ following NISQ feasibility considerations [17]. Single-qubit rotations are applied in parallel and entanglement is restricted to local controlled-Z chains, with one additional inter-patch connection for the two-patch encoding, to minimize depth while capturing patch-level correlations. Kernel construction scales quadratically in the number of training examples because it requires pairwise kernel evaluations, which motivates the controlled dataset size used in this study.

All experiments were executed in simulation on a CPU-only environment. The implementation used Python, standard numerical libraries, scikit-learn for support vector machine-based classification, and a quantum simulation

framework for kernel-fidelity computation. No physical quantum-hardware runs were performed in this study; kernel fidelities were computed under ideal simulation. The shallow eight-qubit, depth-constrained design was chosen to remain compatible with future execution on near-term quantum devices.

IV. RESULTS AND DISCUSSION

This section presents the evaluation protocol, analyzes the structure induced by the proposed quantum kernel, compares Q-Patch with matched classical baselines, and discusses the practical implications and limitations of the reported results.

A. Evaluation Protocol

We evaluate Q-Patch on a balanced subset derived from LJ Speech [8], using 80 training samples and 20 development (dev) samples drawn from a 100-sample set containing 50 bona fide and 50 spoofed utterances. This intentionally compact setup serves two purposes: it makes repeated quantum-kernel simulation computationally manageable and provides a controlled environment for testing whether patch-based quantum embeddings induce meaningful class structure before moving to larger and more diverse spoofing benchmarks.

Unless stated otherwise, all reported metrics are computed on the dev split using standard decision-score evaluation. We report the area under the receiver operating characteristic curve (AUROC) to summarize ranking quality across thresholds and the equal error rate (EER) to characterize the operating point at which false accepts and false rejects are equal. In addition to these scalar measures, we report kernel-similarity statistics and visualize the kernel matrix to examine how Q-Patch organizes bona fide and spoofed samples in the induced feature space.

Because the dev split contains only 20 samples, the reported AUROC and EER should be interpreted as point estimates from a compact feasibility study rather than as statistically definitive performance claims. Accordingly, we treat the observed gains over the baselines as preliminary evidence and leave repeated-split evaluation, uncertainty estimates, and confidence intervals to future work on larger datasets.

B. Quantum Kernel Structure

A central premise of Q-Patch is that a shallow, structured quantum feature map applied to informative time–frequency patches can produce a kernel in which bona fide and spoofed utterances occupy more separable regions. We examine this hypothesis through fidelity-based kernel similarities across within-class and cross-class comparisons.

The results in Table I show three clear patterns. First, self-similarities for identical patches reach 1.000 ± 0.000 , which is consistent with the fidelity definition and confirms numerical stability of the embedding and overlap computation. Second, within-class similarities across different patches are lower than self-similarities, indicating that the feature map does not collapse diverse samples into a single representation. Third, and most importantly, cross-class similarities between bona fide and spoofed samples are lower (0.614 – 0.616) than same-class

TABLE I: Quantum kernel similarity results on the validation dataset

Comparison Type	Similarity Score	Δ from Ideal
Bona fide–Bona fide (same patch)	1.000 ± 0.000	Baseline
Spoof–Spoof (same patch)	1.000 ± 0.000	Baseline
Bona fide–Bona fide (different)	0.675 ± 0.023	–32.5%
Spoof–Spoof (different)	0.622 ± 0.018	–37.8%
Bona fide–Spoof (patch 1)	0.614 ± 0.015	–38.6%
Bona fide–Spoof (patch 2)	0.616 ± 0.017	–38.4%

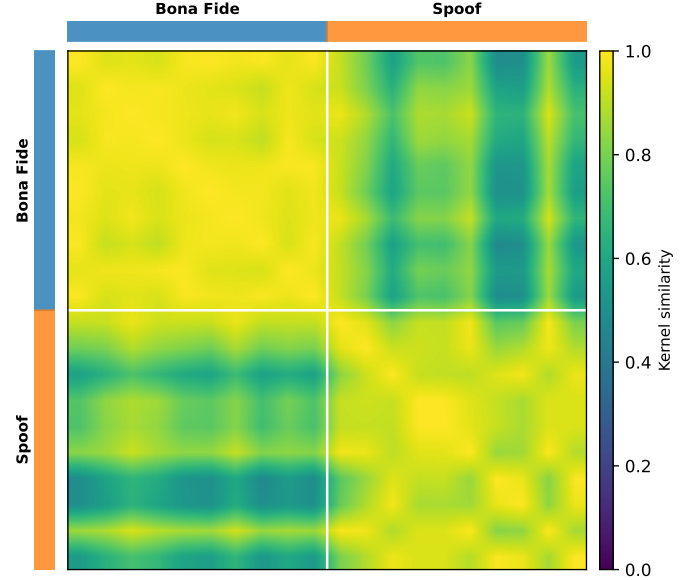


Fig. 4: Quantum kernel similarity matrix on the development set, with samples ordered by class. Brighter within-class blocks and comparatively darker cross-class regions indicate that Q-Patch induces a class-consistent similarity structure between bona fide and spoofed audio samples.

similarities across different patches, suggesting that the induced kernel captures class-relevant structure rather than only generic patch variability.

In Table I, “same patch” refers to self-similarity of an identical patch under the fidelity kernel, whereas “different” refers to patches drawn from different utterances within the same class. A useful way to interpret the table is to compare the drop in similarity caused by changing the patch within a class with the drop caused by changing the class label itself. The 38.4–38.6% reduction observed for bona fide–spoof comparisons matches or exceeds the reduction observed for within-class comparisons across different patches (32.5–37.8%). This indicates that the kernel is responsive to authenticity-related cues beyond ordinary within-class variation. The similarity matrix in Fig. 4 reinforces this conclusion by showing visible clustering consistent with class separation in the induced feature space.

TABLE II: Performance comparison on the validation dataset

Method	AUROC	EER (%)	Complexity
RBF-SVM	0.82	18.2	Support vectors
Tiny CNN	0.85	16.3	98.4K params
Q-Patch	0.87	14.8	8 qubits

C. Comparative Performance

Table II summarizes the classification performance of Q-Patch and the two matched classical baselines. Q-Patch achieves an AUROC of 0.87 and an EER of 14.8%, outperforming both the radial basis function support vector machine (RBF-SVM) trained on the same patch-level descriptors and the compact convolutional neural network (Tiny CNN) trained directly on spectrograms.

The comparison with the RBF-SVM is particularly informative because both methods operate on the same summarized patch features. This makes the RBF-SVM a direct control for assessing the contribution of the proposed quantum feature map and fidelity-based kernel. The improvement over the Tiny CNN further suggests that Q-Patch remains competitive even against a nonlinear model trained on higher-dimensional spectrogram inputs, while maintaining a compact effective quantum footprint of eight qubits and a depth-constrained circuit.

From an application perspective, the lower EER indicates a more favorable trade-off between false accepts and false rejects under the same evaluation protocol. Taken together with the kernel-space analysis, these results support the view that the proposed patch-based quantum embedding is not only expressive, but also practically discriminative in this controlled setting.

D. Interpretation in the Limited-Data Setting

Although the proposed protocol uses only 80 training samples, it still represents a low-data regime relative to typical deep anti-spoofing pipelines. Q-Patch is well suited to such conditions because it combines two desirable properties: aggressive dimensionality reduction through patch summarization and top- k selection, and margin-based learning in kernel space.

Concretely, each patch summary $\mathbf{s} \in \mathbb{R}^4$ is embedded into a quantum state in $\mathcal{H} = \mathbb{C}^{2^n}$, with $n = 4$ qubits per patch block, and similarity is measured through the fidelity kernel

$$k(\mathbf{x}, \mathbf{x}') = |\langle \phi(\mathbf{x}) | \phi(\mathbf{x}') \rangle|^2, \quad (23)$$

where $\phi(\cdot)$ denotes the implicit feature map induced by the Q-Patch circuit. This construction introduces controlled nonlinearity through shallow entanglement while avoiding the large number of trainable parameters typically associated with deep models. In practical terms, the resulting Quantum Support Vector Machine (QSVM) promotes large-margin separation in the induced feature space, which can help reduce overfitting when labeled data are limited.

E. Practical Feasibility and Scalability

Q-Patch is designed with near-term quantum constraints in mind. The two-patch configuration requires an eight-qubit circuit with depth limited to $d \leq 3$ and a minimal entangling pattern consisting of three controlled-Z gates per patch block plus a single inter-patch connection, which is consistent with Noisy Intermediate-Scale Quantum (NISQ) feasibility guidelines [17]. Because single-qubit rotations can be executed in parallel, the circuit remains shallow and comparatively lightweight.

At the same time, the main computational bottleneck lies in kernel construction. For a training set of size N , QSVM requires $\mathcal{O}(N^2)$ kernel evaluations. In simulation, these values are obtained through state overlaps; on hardware, they would need to be estimated through repeated measurements with a shot budget M , creating a trade-off among dataset size, estimation variance, and runtime. This is precisely where the patch-summarization strategy becomes valuable: it limits qubit count and circuit depth while preserving informative structure from the time-frequency representation.

F. Limitations and Scope

The present study is intended as a controlled feasibility analysis, and several limitations should be kept in mind when interpreting the results.

a) *Controlled spoof generation.*: The spoofed samples are generated through additive noise and spectral distortions. These manipulations provide a useful controlled proxy for authenticity variation, but they do not capture the full diversity of real-world attacks, including replay conditions, modern neural text-to-speech systems, and voice conversion artifacts. The reported results should therefore be interpreted as evidence of discriminative potential in a controlled setting rather than as a claim of comprehensive anti-spoof robustness.

b) *Dataset size and statistical uncertainty.*: The dev split contains only 20 samples, which limits the granularity of performance estimation and increases uncertainty in the reported AUROC and EER values. Although the kernel-space trends are consistent, broader evaluation protocols will be needed to establish stronger claims about generalization.

c) *Simulation-to-hardware gap.*: The current study evaluates kernel values under ideal quantum simulation. Real quantum hardware introduces noise, gate errors, decoherence, and finite-shot effects, all of which can perturb fidelity estimates. Although Q-Patch is deliberately designed to remain shallow and hardware-aware [17], additional studies under realistic noise models or on physical devices will be necessary to characterize deployment robustness. For this reason, the present results should be regarded as ideal-simulation feasibility results rather than hardware-level performance estimates.

d) *Sensitivity to preprocessing and patch selection.*: The proposed framework depends on the preprocessing pipeline, the patch size, the top- k criterion, and the selected summary statistics, all of which influence which spectrogram regions are ultimately encoded. A broader ablation study over patch size, k , circuit depth, and entanglement pattern was not feasible

within the present compute and page budget. The reported configuration should therefore be viewed as a principled design point rather than an exhaustive exploration of the design space.

G. Sustainability Considerations

No direct emissions estimate was made for the present study. Since all results are based on small-scale simulation, carbon reporting is best framed here as a recommendation for future larger-scale or hardware-based evaluations rather than as a completed measurement. A practical next step would be to record runtime, energy consumption, and estimated CO₂e during feature extraction, model training, and kernel computation using standard accounting tools. Such reporting would support more transparent comparisons of accuracy–compute trade-offs between classical baselines and quantum-kernel approaches as evaluation scales increase.

V. CONCLUSION

This paper presented *Q-Patch*, a patch-based quantum feature-mapping framework for audio spoofing detection that encodes informative local time–frequency regions into shallow quantum circuits and performs classification through a fidelity-based kernel. Experimental results on a controlled and balanced validation setup showed that *Q-Patch* induces class-consistent kernel structure and improves performance over matched classical baselines, indicating that time–frequency-aware quantum feature maps can provide a useful inductive bias for audio authenticity discrimination in limited-data settings. In addition, the proposed framework remains computationally lightweight by combining a shallow, deterministic quantum embedding with standard kernel-based classification.

The present work should be viewed as a feasibility study under compute and circuit-depth constraints. Future research will focus on evaluation over larger and more diverse spoofing benchmarks, robustness under realistic noise and hardware effects, and scalable kernel-estimation or approximation methods for practical deployment.

REFERENCES

- [1] Z. Khanjani, G. Watson, and V. P. Janeja, “Audio deepfakes: A survey,” *Frontiers in Big Data*, vol. 5, p. 1001063, 2023.
- [2] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, p. 101114, 2020.
- [3] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, “ADD 2022: the first audio deep synthesis detection challenge,” in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9216–9220.
- [4] Y. Ahmadiadli, X.-P. Zhang, and N. Khan, “Beyond identity: A generalizable approach for deepfake audio detection,” *arXiv preprint arXiv:2505.06766*, 2025.
- [5] C. Gao, M. Postiglione, I. Gortner, S. Kraus, and V. Subrahmanian, “Perturbed public voices (P2V): A dataset for robust audio deepfake detection,” *arXiv preprint arXiv:2508.10949*, 2025.
- [6] H. Al-Tairi, A. Javed, T. Khan, and A. K. J. Saudagar, “DeepLASD countermeasure for logical access audio spoofing,” *Scientific Reports*, vol. 15, no. 1, p. 20839, 2025.
- [7] L. Pham, P. Lam, D. Tran, H. Tang, T. Nguyen, A. Schindler, F. Skopik, A. Polonsky, and H. C. Vu, “A comprehensive survey with critical analysis for deepfake speech detection,” *Computer Science Review*, vol. 57, p. 100757, 2025.
- [8] K. Ito and L. Johnson, “The LJ speech dataset,” <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [9] Z. Khanjani, G. Watson, and V. P. Janeja, “How deep are the fakes? focusing on audio deepfake: A survey,” *arXiv preprint arXiv:2111.14203*, 2021.
- [10] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with RawNet2,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.
- [11] L. Al Amin, M. I. Hossain, T. T. Nguyen, T. Jahan, M. Islam, and F. Quader, “Uncovering critical features for deepfake detection through the lottery ticket hypothesis,” in *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2025, pp. 6109–6115.
- [12] R. Hossain, L. E. Ali, and K. S. N. Ripon, “Risk-controlled multimodal emotion coaching for autism support using self-supervised vision and speech encoders,” in *2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2025, pp. 1–7.
- [13] J. Schnabel and M. Roth, “Quantum kernel methods under scrutiny: a benchmarking study,” *Quantum Machine Intelligence*, vol. 7, no. 1, p. 58, 2025.
- [14] S. Egginger, A. Sakhnenko, and J. M. Lorenz, “A hyperparameter study for quantum kernel methods,” *Quantum Machine Intelligence*, vol. 6, no. 2, p. 44, 2024.
- [15] N. Innan, M. A.-Z. Khan, B. Panda, and M. Bennai, “Enhancing quantum support vector machines through variational kernel training,” *arXiv preprint arXiv:2305.06063*, 2023.
- [16] D. Beaulieu, D. Miracle, A. Pham, and W. Scherr, “Quantum kernel for image classification of real world manufacturing defects,” *arXiv preprint arXiv:2212.08693*, 2022.
- [17] H. Tran, B. Kashyap, and P. N. Pathirana, “Quantum approaches for dysphonia assessment in small speech datasets,” *arXiv preprint arXiv:2502.08968*, 2025.
- [18] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke *et al.*, “Noisy intermediate-scale quantum algorithms,” *Reviews of Modern Physics*, vol. 94, no. 1, p. 015004, 2022.
- [19] M. Schuld and N. Killoran, “Is quantum advantage the right goal for quantum machine learning?” *PRX Quantum*, vol. 3, no. 3, p. 030101, 2022.
- [20] L. Al Amin and V. P. Janeja, “Reliable audio deepfake detection in variable conditions via quantum-kernel SVMs,” in *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2025, pp. 1395–1403.