

A Comparative Study of Quantum and Hybrid Models for Audio Deepfake Detection

Rahma Alam Samiha

World University of Bangladesh

Md. Tasbirul Alom

World University of Bangladesh

Ajmery Sultana

Algoma University

Rubaiyat Islam

`islam.swe@diu.edu.bd`

Daffodil International University

Research Article

Keywords: quantum machine learning, deepfake speech, hybrid quantum, deepfake detection

Posted Date: June 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9320777/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Comparative Study of Quantum and Hybrid Models for Audio Deepfake Detection

Rahma Alam Samiha¹, Md. Tasbirul Alom¹, Ajmery Sultana²,
Rubaiyat Islam^{3*}

¹Dept. of Computer Science & Eng., World University of Bangladesh,
Dhaka, Bangladesh.

²Dept. of Computer Science & Mathematics, Algoma University,
Brampton, Ontario, Canada.

³Dept. of Software Eng., Daffodil International University, Dhaka,
Bangladesh.

*Corresponding author(s). E-mail(s): islam.swe@diu.edu.bd;

Contributing authors: 0320533345@student.wub.edu.bd;

0320533338@student.wub.edu.bd; ajmery.sultana@algonau.ca;

Abstract

The recent advancement in deepfake creation has become a critical issue in the domain of digital security. These synthetic voices are difficult to identify from real voices. Currently, detection frameworks and datasets for low-resource languages like Bangla are not widely available. This is a critical research gap. In order to fill the research gap, a novel Bangla deepfake audio dataset is introduced in this paper. In addition, the paper proposes the application of quantum machine learning for the detection of deepfakes. In the paper, four architectures have been investigated. These include, Quantum Support Vector Machine (QSVM), Quantum Convolutional Neural Network (QCNN), Quantum Long Short-Term Memory (QLSTM), and the combination of QCNN and QSVM. The hybrid system has achieved an accuracy of 99.5%. It has surpassed the accuracy and convergence rate of the standalone models with the addition of QCNN's hierarchical feature encoding and QSVM's margin-based decision boundary. The experimental results demonstrate the performance of the hybrid quantum systems in comparison to their standalone counterparts. This is the first time Quantum and Hybrid-Quantum Learning have been formally applied to the problem of deepfake speech detection in a low-resource language. These results provide significant insight into the potential of using Hybrid Quantum systems as a powerful tool for multimedia forensic analysis.

Keywords: quantum machine learning, deepfake speech, hybrid quantum, deepfake detection

1 Introduction

Today deepfakes pose a critical threat to digital security, privacy, and information integrity. Deepfakes are artificially created media that are designed with the help of advanced algorithms. They can mimic human faces, voices, and emotions with great precision. There are different types of deepfakes. While deepfakes exist in many forms, audio deepfakes are equally concerning. Existing voice-cloning technology can easily copy a person’s tone, pitch, and rhythm of speech based on just a few seconds of speaking. This allows the creation of deepfake recordings that are very similar to the authentic ones. These artificial voices can be used to spread false information, impersonate people, commit fraud, or bypass voice verification systems [1–3]. Therefore, these technologies have become more widely available on open-source platforms and commercial software. For these reasons, identifying real and fake audio has become a technical concern, as well as a social obligation.

In response to this threat, many researchers have turned towards different Machine Learning (ML) and Deep Learning (DL) techniques to bring solutions. These techniques have significantly enhanced deepfake detection. They can detect small variations between real and fake audio, which is difficult for humans to detect. Traditional approaches followed hand-designed acoustic features, such as Mel Frequency Cepstral Coefficient (MFCC). These features represent the spectral and tonal content of speech. Later developments in deep learning introduced more powerful models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. Our previous work experimented with the quality of detection over different dimensions of CNN architectures[4]. These technologies can learn rich representations directly from raw spectrograms and waveforms [5–7]. They have produced high accuracy and good performance on many benchmark datasets. However, their generalization ability remains limited. Moreover, most ML and DL models are inconsistent across different datasets, spoofing attacks, and speaker accents. Additionally, the majority of research has been conducted using resource-intensive languages like English. However, low-resource languages, such as Bengali, remain relatively unexplored [8–10]. This disparity limits the overall fairness and integrity of existing deepfake detection techniques.

In recent years, Quantum Machine Learning (QML) has been seen as a growing field. This field combines the technology of quantum computing with the power of ML. QML is said to have the advantages of both quantum computing and traditional ML. This is achieved by applying the unique properties of quantum mechanics to process and represent information. We know, in classical computation, information is represented in bits that are either 0 or 1. However, quantum computation depends on qubits. This means, unlike classical bits, they can represent multiple values at once. This is achieved by using superposition and entanglement. When qubits are entangled, they model complex relationships of data in exponentially larger spaces

than classical systems. Such spaces are known as Hilbert spaces [11]. This unique property enables QML models to mimic complex data correlations. Therefore, this should offer potential advantages in training efficiency, generalization, and robustness against noise or adversarial attacks. [12, 13].

Besides being a growing research field, QML is already promising across a broad spectrum of applications. Quantum algorithms such as Quantum Support Vector Machines and Quantum Convolutional Neural Networks have performed well in image classification, speech signal processing, and secure communication [14, 15]. Apart from the enhancement of performance, QML also offers potential solutions to classical ML challenges such as overfitting, scalability, and computational inefficiency by using quantum parallelism to handle data in high-dimensional spaces [16]. In addition, due to its hybrid nature in which quantum circuits are crossed with classical optimization, QML may be portable to data-hungry and dynamic problem spaces [17]. Research concerning quantum-aided 6G networks and intelligent communication protocols also portrays QML’s ability to guarantee energy-efficient processing, adapting to security threats, and cybersecurity [18]. Despite being in its infancy, QML is a growing and much-needed area of research that must be experimented with continuously to realize its full potential.

Motivated by this, we examine QML on a Bengali deepfake audio. This language is largely overlooked in AI research when viewed from a global perspective. This research provides a constructive evaluation of the applicability of Quantum ML on a low-resource language from a speech perspective. Ultimately, we aimed to make a candid evaluation of what is capable within the realm of quantum models at present. The major contributions to this research are:

- For our research, we used our self-created “Bengali Real and Deepfake Audio Dataset for Deepfake Detection” [19], with a motive to extend the research on deepfakes even in a low-resource environment.
- We implemented standalone QSVM, QCNN, and QLSTM models, and tested them on different cases. These models were compared and evaluated based on varying numbers of qubits, encoding types, and convergence rates. This allowed us to check the efficiency of the models under different conditions.
- We proposed a hybrid architecture of the two most successful models, QCNN and QSVM. We examined whether a combination of models is more synergistic than the standalone models.
- In contrast to research that only focused on accuracy, we assessed training speed, execution speed, and performance metrics such as the F1 score to determine the potential of QML. Our proposed QCNN→QSVM hybrid converged faster than QSVM in just 22 epochs. The execution time of the hybrid is also faster, taking 18.2 seconds per round compared to 21.3 seconds for QSVM.

The rest of the paper is structured as follows. Section II presents an overview of some related works on audio deepfakes and ML on quantum computers. Section III introduces the research methodology, including data preparation, model definition, and the combination method. Section IV presents the experiment and the settings for

defining several models, and the results obtained on accuracy and efficiency. Lastly, Section V concludes the paper by stating some findings and limitations.

2 Related Work

Deepfake audio detection has become a dynamic research area, as identifying real versus synthetic voices is crucial for sustaining digital trust. Most classical deep learning models exhibit strong performance in this area. However, they are mostly based on large data sets, are computationally expensive, and can hardly generalize well under unseen or adversarial settings. Recently, QML has emerged as a potential solution that enables dealing with small and noisy datasets by using the powers of superposition and entanglement. This section reviews prior work in three domains. These include classical deepfake detection, quantum models applied to audio and signal processing, and quantum or hybrid approaches for deepfake detection, followed by an overall discussion of the identified research gaps and how our work addresses them. A summarized view can be found in the Table 1.

2.1 Classical Audio Deepfake Detection

Several works have established strong deep learning benchmarks for detecting audio deepfakes. As mentioned earlier, our previous work delved into the deep learning (DL) framework of CNN architectures. The study implemented different CNN architectures based on their dimensions [4]. It achieved a 98.24% accuracy rate with the 2-dimensional (2D) CNN architecture. Another study, Pham et al. [20], used an ensemble spectrogram-based model and reported an EER of 0.03 on the ASVspoof2019 dataset. This showed that feature fusion is powerful for detecting spoofed speech. However, their approach used heavy computation and failed to evaluate unseen data. Kanwal et al. [21] introduced attention mechanisms into a VGGish-based architecture that achieved EERs of 0.52% for Physical Access and 0.07% for Logical Access. While these were promising results, their performance greatly depended on the dataset characteristics.

Yadav et al. [22] have proposed a “Disentangled Spectrogram Variational Auto Encoder” (DSVAE). This could separate real and fake latent features to get more than 98% accuracy. However, their model was validated on very limited datasets, thus restricting the generalizability. Xie et al. [23] proposed the multi-dataset named “FakeSound” benchmark and obtained better performance compared to the best state-of-the-art systems. Although localization and dataset standardization remain open issues.

Other works, such as Kreuk et al. [24] and Wolter et al. [25], investigated vocoder artifacts and lightweight architectures, respectively, for better generalization. In general, classical models achieve high accuracy with great data scale and computation, but usually do not generalize well to other spoofing methods, accents, or recording conditions.

Table 1: Key Existing Related Work in Audio Deepfake Detection and QML

Paper	Type	Models	Results	Limitations
Pandey & Rudra [26]	Deepfake (Audio)	QSVM, QNN, Hybrid CNN-QNN	QSVM: 90%, QNN: 70%, Hybrid: 98.8%	Pure quantum weak; simulator limits
Lin et al. [27]	Deepfake (Audio)	QT-CNN	Maintains accuracy with $\sim 70\%$ fewer parameters	Scalability issues with qubits
Quantum-Powered Deepfake Detection [28]	Deepfake (Audio/Video)	QSVM, QNN, QFT, QAA	95.2% accuracy, 95% F1, 30% faster	Needs quantum hardware; complex integration
Pham et al. [20]	Deepfake (Audio)	Spectrogram + Ensembles	EER = 0.03 (ASVspoof2019)	High computation; unclear generalization
Kanwal et al. [21]	Deepfake (Audio)	VGGish + Attention	EER = 0.52% (PA), 0.07% (LA)	Dataset-specific; limited scope
Yadav et al. [22]	Deepfake (Audio)	DSVAE	>98% accuracy	Tested on limited datasets
Xie et al. [23]	Deepfake (Audio)	FakeSound Benchmark	Outperforms prior SOTA	New dataset; localization challenge

2.2 Quantum Methods for Audio and Signal Processing

Recently, quantum-based learning methods have shown promising results regarding audio and signal-related tasks. Chaharlang et al. [29] applied Quantum Haar Wavelet Transform with Quantum Support Vector Machines for audio steganalysis. Their results showed that quantum learning performed better in the frequency domain. Although the study only focused on steganography and not deepfakes.

Thakar et al. [30] proposed a hybrid quantum-classical convolutional neural network for the classification of audio spectrograms. It achieved an accuracy of 95.04%, superior to that of classical CNNs, which was 92.88%. Consequently, it also reduced overfitting. Hasmitha Krishna et al. [31] employed the quantum kernel method to classify speech using merely 4–13 qubits. It achieved an accuracy of 87.32%, demonstrating the efficiency of small-scale quantum models. The authors of [32] proposed Hybrid Quantum Neural Networks for sound tasks in healthcare, such as cough detection. While promising, they demonstrated high sensitivity to quantum noise and extended training times. Wu et al. [33] applied quantum variational classifiers to radio signals, demonstrating greater robustness against adversarial attacks than CNNs.

Further studies, such as decentralized QCNNs for speech recognition [34], staggered hybrid CNNs (St-HQCNN) for environmental sound detection [35], and PQC-CNNs for emotion recognition [36], consistently report accuracies above 95%. These works

collectively highlight QML’s ability to handle small data, reduce overfitting, and improve robustness. However, none directly target audio deepfake detection.

2.3 Quantum and Hybrid Approaches for Deepfake Detection

Research that combines QML and deepfake detection is still in its infancy. The work by Katı et al. [37] used Quantum Transfer Learning (QTL) with vision transformers. It achieved 90% accuracy and 0.94 Area Under the Receiver Operating Characteristic (AUROC). However, their model was restricted to visual data and was tested only in simulation.

Pandey and Rudra [26] studied the problem of deepfake audio detection with purely quantum and hybrid models. While QSVM and QNN achieved relatively modest performances at 90% and 70%, their CNN-QNN hybrid achieved 98.81% accuracy, very close to the classical CNN at 99.19%. Thus, it showed that hybrids can effectively balance quantum and classical advantages.

Lin et al. [27] proposed a Quantum-Trained CNN where quantum layers could optimize the CNN parameters, reducing the total number of parameters by around 70% without compromising much accuracy. A more general framework of “Quantum-Powered Deepfake Detection” [28] combined QSVMs, QNNs, Quantum Fourier Transform, and Quantum Amplitude Amplification, which realized an accuracy of 95.2% and 30% faster inference, but with a more complex integration of hardware. Likewise, the ResNet + QLSTM hybrid [38] reduced the parameters by 95% while maintaining high accuracy on Celeb-DF and FaceForensics++.

Although these works illustrate that QML can improve speed, reduce parameters, and preserve accuracy, most of the works are limited to video or general audio tasks without comparative scenario analysis, and normally rely on simulators rather than physical quantum systems.

From these studies, several insights and gaps become clear. QML models show higher performance compared to classical models in conditions where the data is scarce or noisy [31, 39]. They are more robust to adversarial manipulation [33]. Hybrid quantum-classical systems have shown good results when overfitting [30] and better computational efficiency [26–28, 38]. However, purely quantum models are still bound by simulators and limited support for quantum hardware. Hence, they are lagging behind classical architectures [26]. More importantly, very few research works have applied QML specifically to audio deepfake detection. Moreover, an absence of structured experimental comparisons, low-resource language evaluation, and multi-scenario analysis is noticed. This points out the need for systematic experiments that provide insight into how QML behaves across different scales of data, qubit configurations, and training dynamics.

This section describes the methodology adopted to design and evaluate our QML models for Bangla audio deepfake detection. The workflow has five components: dataset preparation, preprocessing, quantum feature encoding, model development, and comparative analysis. Four architectures were explored: QSVM, QCNN, QLSTM, and a hybrid QCNN→QSVM configuration. Further, hybrid paradigms, the overall data pipeline, and the principles underlying each model are discussed in detail.

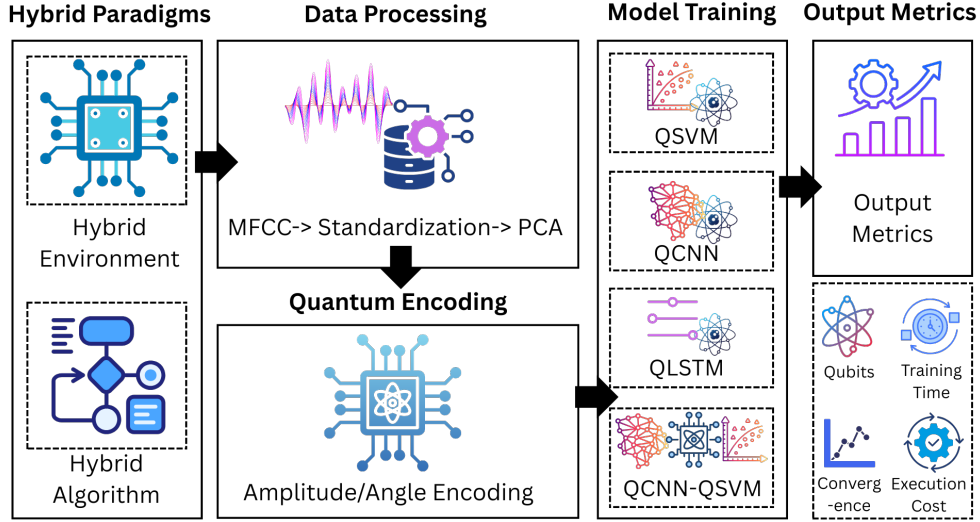


Fig. 1: Proposed Research Framework for QML Model Evaluation

3 Methodology

This section describes the methodology adopted to design and evaluate our QML models for Bangla audio deepfake detection. The workflow has five components: dataset preparation, preprocessing, quantum feature encoding, model development, and comparative analysis. Four architectures were explored: QSVM, QCNN, QLSTM, and a hybrid QCNN→QSVM configuration. Further, hybrid paradigms, the overall data pipeline, and the principles underlying each model are discussed in detail.

3.1 Hybrid Paradigms

In this work, two hybrid complementary concepts were adopted.

Hybrid environment: The work of classical and quantum computations is integrated into one workflow. Each of these handles tasks that are best suited to its strengths. The classical subsystem was in charge of data pre-processing (standardization, normalization, and PCA). It utilized a variety of gradient-based optimizers, such as Adam or Optax, for effective parameter updates. The quantum subsystem handled feature encoding, circuit evolution, and measurement. Quantum outputs during training pass through classical routines to compute losses. The ensuing gradients then feed back to adjust the quantum gate parameters. This interaction enables shallow, stable, and computationally efficient circuits even with limited qubits.

Hybrid algorithms: This hybrid concept concatenates diverse quantum models in a serial manner such that the output from one model becomes the input for another model. Our work applies this approach by the QCNN→QSVM arrangement. Here, the QCNN acts as a quantum feature extractor entailing convolution and pooling, and QSVM opts for margin-based classification using a quantum kernel. Isolating

feature extraction and decision-making enhances the interpretability and accelerates the convergence of the solution as opposed to a solo quantum model.

3.2 Pipeline Overview

This experimental pipeline consists of three main steps as represented by Fig. 1:

Stage 1 — Data Preparation. Raw audio files were transformed into Mel-Frequency Cepstral Coefficient tensors, standardized, and reduced in dimension using PCA to align feature sizes with qubit capacities.

Stage 2 — Quantum Encoding. The reduced features were encoded into quantum states through the use of Amplitude Encoding (AE) or Angle Encoding (AngE). AE embeds normalized classical vectors into amplitude coefficients, while AngE encodes features as rotation angles on individual qubits.

Stage 3 — Model Training and Evaluation. Data in coded format were analyzed using either of the four models: QSVM, QCNN, QLSTM, or the hybrid QCNN→QSVM. All of these models provided binary predictions as output: bona fide and fake, whose performance could be evaluated based on accuracy, AUROC, EER, F1-score, and computational efficiency. This structure ensures consistency across experiments.

3.3 Dataset

The dataset used for all the experiments consisted of 8,000 audio samples with 40 MFCC features over 281 frames [19]. This data was divided into training, validation, and test sets and had a balanced distribution of real and deepfake audio. The overall number of samples was kept the same across all the models, while the actual number of the train and test samples is different depending on the model and the exact experimental setting.

In the case of the QSVM model, we used 420 training samples and 140 test samples for 4 qubits with amplitude encoding. Then, 840 training samples and 280 test samples for 6 qubits with amplitude encoding. These were the splits that had to be used in order to test the generalization of the model for different numbers of qubits and types of encoding. Therefore, QCNN was trained on a dataset that was larger in comparison to QLSTM. The number of training samples for QCNN was 700, and the number of test samples was 300. With an increase in training samples, the features extracted from training samples by QCNN improved, increasing its performance. QLSTM was trained with much fewer samples in its dataset, utilizing 50 training samples and only 20 test samples. This is indicative of the limitation brought about by the number of available training samples and the model complexity itself. The smaller size of the dataset reduced the performance of the model.

The number of training samples and test samples for the QCNN→QSVM Hybrid model is 600 and 200, respectively. This model merges the power of feature extraction by QCNN with the classification power brought by QSVM. The size of the dataset used has to be in a balance between efficiency and accuracy, balancing sufficient training without a high computational cost.

Standardization and PCA were applied to the dataset for preprocessing because they are very helpful in improving the performance of quantum models, especially when feature encodings were applied. These varying dataset sizes reflect the model configurations tested during the experiment.

3.4 Data Preprocessing

Let $X \in \mathbb{R}^{N \times n_{\text{mfcc}} \times n_{\text{frames}}}$ represent the MFCC tensor and $y \in \{0, 1\}$ the labels.

For non-sequential models (QSVM, QCNN), samples were flattened, standardized, and reduced via PCA:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}, \quad \mathbf{z} = \mathbf{W}_d^\top \tilde{\mathbf{x}}, \quad (1)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the training mean and standard deviation, and $\mathbf{W}_d$ contains the top d principal components. The feature vector was padded or truncated to match the qubit dimension:

$$\mathbf{z}^{(d)} = \begin{cases} [\mathbf{z}; \mathbf{0}], & \dim(\mathbf{z}) < d, \\ \mathbf{z}_{1:d}, & \text{otherwise.} \end{cases} \quad (2)$$

For QLSTM, temporal order was preserved. Each sample was transposed to $(n_{\text{frames}}, n_{\text{mfcc}})$, standardized per frame, reduced to F features, and aligned to a fixed sequence length T :

$$\mathbf{S} = (\mathbf{s}_1, \dots, \mathbf{s}_T) \in \mathbb{R}^{T \times F}, \quad \mathbf{s}_t = \mathbf{W}_F^\top \tilde{\mathbf{x}}_t. \quad (3)$$

All scalars and PCA mappings were fit on the training set and reused during validation and testing.

3.5 Quantum Encodings

Two encoding strategies were adopted to map classical data to quantum states.

Amplitude Encoding (AE). When $d = 2^n$, features were normalized and loaded as quantum state amplitudes:

$$\hat{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \quad |\psi_{\text{AE}}(\mathbf{z})\rangle = \sum_{i=0}^{2^n-1} \hat{z}_i |i\rangle. \quad (4)$$

Angle Encoding (AngE). For $d = n$, features were mapped as qubit rotation angles:

$$|\psi_{\text{AngE}}(\mathbf{z})\rangle = \left(\prod_{i=1}^n R_Y(z_i) \right) |0\rangle^{\otimes n}. \quad (5)$$

AE was applied when PCA output matched 2^n , while AngE was used for shorter or sequential data (e.g., QLSTM).

3.6 Model Descriptions and Implementation

In this section, we describe the technical architecture of the quantum models one by one that are being investigated. To have fine control over qubit manipulations and a

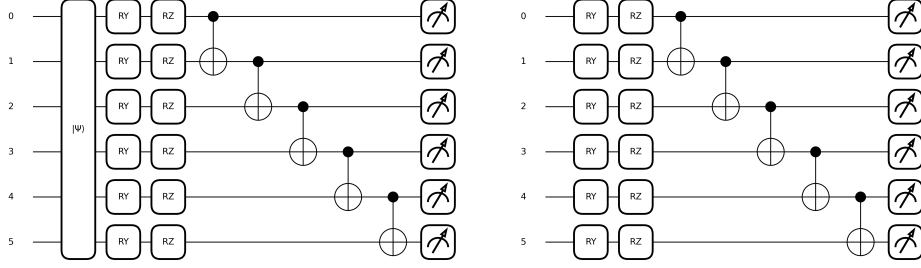


Fig. 2: QSVN feature maps: amplitude embedding (left) and angle embedding (right).

seamless implementation with classical data representations, all the quantum models are implemented within the PennyLane environment. The technical architecture of QSVN, QCNN, as well as the QLSTM is provided in Fig. 2, Fig. 3, and Fig. 4, respectively. In addition, the architecture of our proposed Hybrid Model is displayed in Fig. 5.

3.6.1 Quantum Support Vector Machine (QSVN)

QSVN extends classical SVM by leveraging a quantum feature map to project data into a high-dimensional Hilbert space. The separation between classes is achieved using a quantum kernel that measures the overlap between quantum states. This allows QSVNs to represent complex, nonlinear decision boundaries using fewer parameters.

Given an input vector $\mathbf{z}$, the feature map $U_\phi(\mathbf{z})$ prepares a quantum state:

$$|\psi(\mathbf{z})\rangle = U_\phi(\mathbf{z}) |0\rangle^{\otimes n}. \quad (6)$$

The kernel between two states is:

$$K_{ij} = |\langle\psi(\mathbf{x}_i)|\psi(\mathbf{x}_j)\rangle|^2. \quad (7)$$

Optimization follows the standard SVM dual form, while classification uses the decision rule:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\right). \quad (8)$$

QSVN was implemented in PennyLane using AE and AngE on $n = 4$ and $n = 6$ qubits. CNOT-based entangling layers modeled inter-qubit correlations, and kernel overlaps were estimated via *SWAP* tests. The dual optimization problem was solved with classical solvers. We used 420/140 and 840/280 splits for 4- and 6-qubit configurations, respectively.

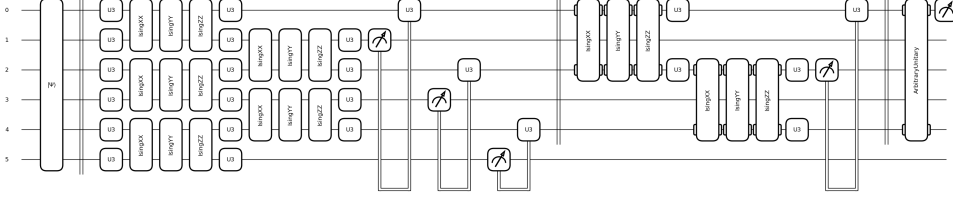


Fig. 3: QCNN circuit showing convolution, pooling, and dense layers.

3.6.2 Quantum Convolutional Neural Network (QCNN)

QCNNs are inspired by classical CNNs, but operate through quantum gates. They apply local parameterized unitaries (convolutions) followed by pooling operations (partial trace) to progressively reduce qubits while learning hierarchical features. This structure allows QCNNs to extract spatial or local information efficiently within quantum circuits.

A convolutional unitary for layer ℓ is:

$$U_{\text{conv}}^{(\ell)} = U_{\text{sq}}^{(\ell)} U_{\text{Ising}}^{(\ell)} U_{\text{sq}}^{(\ell)\dagger}, \quad (9)$$

and pooling reduces qubits via

$$\rho_{\ell+1} = \text{Tr}_{\mathcal{P}_\ell} \left(U_{\text{conv}}^{(\ell)} \rho_\ell U_{\text{conv}}^{(\ell)\dagger} \right) \quad (10)$$

After the final layer, the classification probability is obtained as

$$p_{\theta}(y=1|\mathbf{z}) = \frac{1 - \langle Z_{q^*} \rangle_{\rho_L}}{2} \quad (11)$$

Our QCNN used 6 qubits with amplitude encoding. Each block alternated R_X , R_Y , and Ising gates, and pooling progressively reduced active qubits to limit depth. Training minimized binary cross-entropy loss with gradients computed via the parameter-shift rule. We used 700/300 train/test samples and the Optax/Adam optimizer.

3.6.3 Quantum Long Short-Term Memory (QLSTM)

QLSTM extends the classical LSTM by replacing certain linear transformations with quantum circuits. Quantum gates produce activation values through expectation measurements, which are then used in classical recurrent updates. This hybrid mechanism aims to capture temporal dependencies more compactly using quantum entanglement and superposition.

For each time step t , frame $\mathbf{s}_t$ is encoded into a quantum state, producing expectation-based gate activations:

$$\tilde{g}_t^{(k)} = \langle Z_k \rangle_t, \quad g_t^{(k)} = \sigma(a_k \tilde{g}_t^{(k)} + b_k), \quad (12)$$

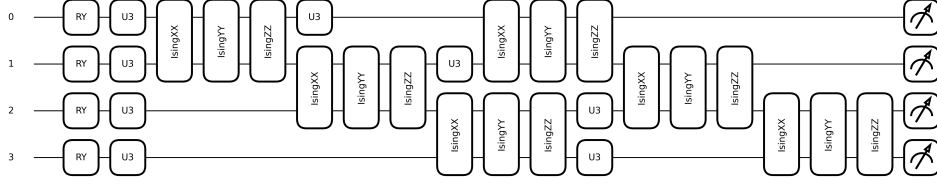


Fig. 4: QLSTM circuit implementing quantum analogues of LSTM gates.

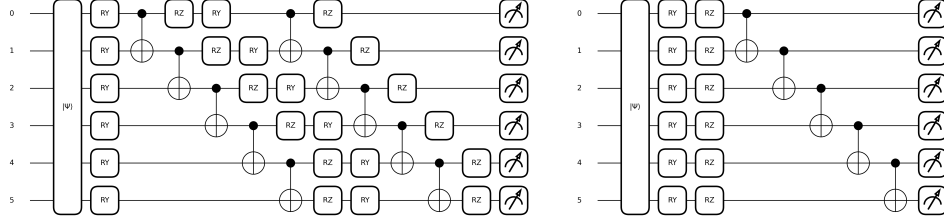


Fig. 5: QCNN→QSVM hybrid: QCNN feature extraction (left) followed by QSVM kernel classification (right).

with cell and hidden state updates:

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \quad (13)$$

The final probability is:

$$p(y=1|\mathbf{S}) = \sigma(\mathbf{w}^\top \mathbf{h}_T + b). \quad (14)$$

QLSTM was implemented with a sequence length of $T = 8$ and a feature dimension of $F = 4$. Each frame was angle-encoded on F qubits using R_Y/R_Z rotations with CNOT entanglers. Training used backpropagation through time (BPTT) and parameter-shift rules. The dataset included 50 training and 20 test samples.

3.6.4 Hybrid QCNN→QSVM

The hybrid QCNN→QSVM model merges QCNN's representational strength with QSVM's classification capability. QCNN extracts compact feature embeddings, which QSVM then classifies using a quantum kernel. This division improves generalization and reduces overfitting.

Feature extraction:

$$\mathbf{h}(\mathbf{z}; \boldsymbol{\theta}) = [\langle Z_{q_1} \rangle_{\rho_L}, \dots, \langle Z_{q_m} \rangle_{\rho_L}]^\top, \quad (15)$$

and hybrid kernel:

$$k_{\text{hyb}}(\mathbf{h}, \mathbf{h}') = |\langle \psi(\mathbf{h}) | \psi(\mathbf{h}') \rangle|^2. \quad (16)$$

Training occurred in two stages: (1) the QCNN was trained to extract discriminative features, and (2) those features were fed into a QSVM trained using the hybrid kernel. This configuration used 6 qubits with 600 training and 200 test samples.

The following section dives into the final results obtained and a complete comparative analysis of our models’ performances.

4 Results and Analysis

This section analyses the performance of four QML models for Bangla audio deep-fake detection: QSVM, QCNN, QLSTM, and the hybrid model QCNN→QSVM. Each model has been investigated in detail for accuracy, F1-score, qubit configuration, encoding method, convergence rate, training efficiency, and computational overhead. The analysis not only compares numerical performance but also interprets the influence of different parameters and design choices on the learning behavior of each architecture.

4.1 Overall Experimental Setup

All experiments were executed in the PennyLane simulation environment on a classical CPU backend. The input audio data was first transformed into Mel-Frequency Cepstral Coefficients, standardized, and then reduced in dimension via Principal Component Analysis. Pre-processed features were then encoded to quantum states through Amplitude Encoding (AE) or Angle Encoding (AngE), depending on the model type and feature length. Generally, AE is used for fixed-length features because of its compact representation of states, while AngE is applied in sequential models like QLSTM in order not to lose temporal information. Before training the full suite of models, QSVM served as a baseline for parameter tuning. Several combinations of optimizers, learning rates, batch sizes, and stopping criteria were explored. From here, the best configuration ensuring maximum stability and accuracy for QSVM was then uniformly ported to QCNN, QLSTM, and the hybrid QCNN→QSVM. This was to avoid having comparative differences among the models caused by inconsistencies in parameters. All reported metrics, including accuracy, F1-score, time of training per epoch, time of execution per batch, and convergence epochs, were measured under identical computational resources for fairness.

4.2 Baseline QSVM Analysis

The QSVM model set the baseline against which further architectures would be compared. Two setups were explored: 4 and 6 qubits with Amplitude Encoding to test performance against representational capability. Test results are shown in Table 2, where it is seen that increasing the number of qubits significantly improved performance. The 6-qubit configuration reached an accuracy of 92.0% and an F1-score of 0.920 after converging at 35 epochs. However, the 4-qubit model reached 89.3% accuracy. Average time to train one epoch was 21.3 seconds, whereas execution per batch took 0.18 seconds.

Table 2: QSVM Baseline Performance under Different Qubit Configurations

Qubits	Encoding	Accuracy (%)	F1-Score	Convergence
4	Amplitude	89.3	0.891	38
6	Amplitude	92.0	0.920	35

4.3 Analysis of QLSTM

The Quantum Long Short-Term Memory (QLSTM) network was trained on 4 qubits with Angle Encoding. From Table 3, it can be noticed that the accuracy and F1-score for the QLSTM network were 40.0% and 0.401, respectively. The number of epochs required for convergence was 52. The average training time per epoch took 32.5 seconds, and the execution time per batch took 0.27 seconds. Although it is theoretical that QLSTM can maintain temporal information, it performed poorly compared to other architectures. The fact that it is a sequential RNN and there were fewer qubits meant that there could be high variances and irregular learning rates. Although Angle Encoding maintained ordering, it could not maintain larger temporal ordering due to limitations within the restricted Hilbert space. The higher training time and slow convergence rates were due to higher computational costs due to sequential gate operations. It can be seen that QLSTMs might perform better with larger quantum circuits.

Table 3: QLSTM Performance Metrics

Metric	Value
Accuracy (%)	40.0
F1-Score	0.401
Training Time / Epoch (s)	32.5
Execution Time / Batch (s)	0.27
Convergence (Epochs)	52
Qubits	4
Encoding Method	Angle Encoding (AngE)

4.4 Analysis of QCNN

In the work, QCNN with 6 qubits and Amplitude Encoding is implemented. According to Table 4, it reached an accuracy of 85.5% and F1-score of 0.854 after convergence at 28 epochs. The average training time was 15.7 sec/epoch, 0.12 sec/execution per batch. As shown in Fig. 6, there are no oscillations in loss and accuracy, and their convergence is smooth and monotonic. Due to the parameterized convolutional and pooling layers, QCNN efficiently captured the local spectral features. Hierarchical learning reduces redundant features while keeping discriminative information in turn. Even though QCNN needs fewer computational resources compared with QSVM, the direct classification accuracy of QCNN is slightly lower due to output probabilities compression by measurement-based readouts, limiting its resolution. However, this

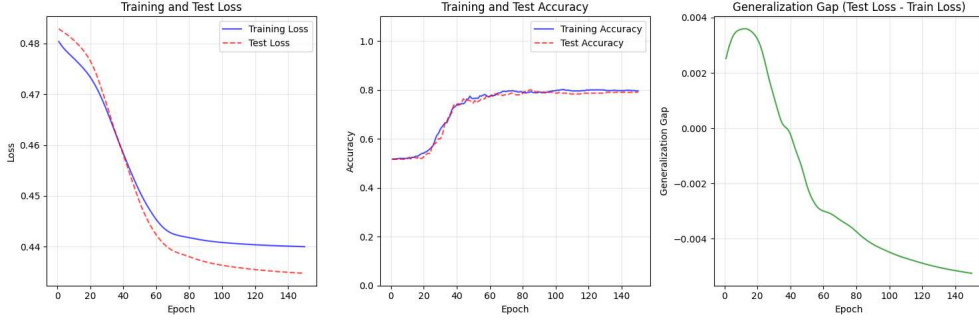


Fig. 6: Training and validation loss and accuracy curves for QCNN. The steady improvement indicates stable learning and effective feature extraction.

makes it highly efficient and stable as a quantum feature extractor for hybrid higher-order systems.

Table 4: QCNN Performance Metrics

Metric	Value
Accuracy (%)	85.5
F1-Score	0.854
Training Time / Epoch (s)	15.7
Execution Time / Batch (s)	0.12
Convergence (Epochs)	28
Qubits	6
Encoding Method	Amplitude Encoding (AE)

4.5 Analysis of Hybrid QCNN→QSVM

The hybrid model QCNN→QSVM comprised quantum convolutional learning for feature extraction and a kernel method for classification. Its performance, as recorded in Table 5, was better than that of all other models as it reached an accuracy and F1-score of 99.5% and 0.995, respectively. Also, it took 22 epochs with 18.2 seconds per epoch and 0.15 seconds per batch. The convergence shown in Fig. 7 represents faster convergence and lower loss value variance. Also, feature embeddings obtained via QCNN were effectively compressed and embedded within quantum, and these were then properly classified using a large-margin quantum kernel. Moreover, based on these serial steps, embeddings were very well-separated with low chances of overfitting. Hence, there was near-perfect accuracy. It had easier computation and accuracy as it took faster convergence compared to all models, but it consumed a larger total number of qubits effectively, because it took 12 qubits.

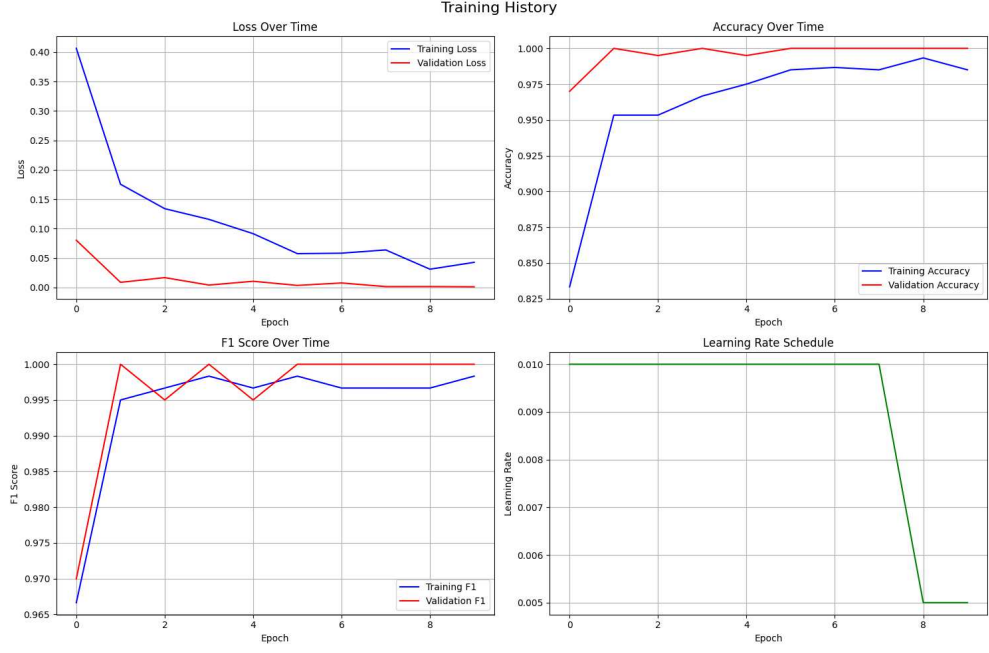


Fig. 7: Training and validation accuracy and loss for the QCNN→QSVM hybrid. Rapid and consistent convergence indicates strong generalization and efficient learning.

Table 5: QCNN→QSVM Hybrid Model Performance Metrics

Metric	Value
Accuracy (%)	99.5
F1-Score	0.995
Training Time / Epoch (s)	18.2
Execution Time / Batch (s)	0.15
Convergence (Epochs)	22
Qubits	12
Encoding Method	Amplitude Encoding (AE)

4.6 Comparative and Performance Analysis

For a better comparison of trade-offs between the detection power and computational cost, we have visualized the performance of all four models in a “Radar Chart” as shown in Fig. 8. Because the metrics are expressed in different units, we normalized them to a scale $[0, 1]$, with the outermost edge corresponding to the best performance. Notably, we normalized cost metrics (Training Time, Qubit Count, Epochs) with the following formula: $x'_{ij} = (\max(X_j) - x_{ij}) / (\max(X_j) - \min(X_j))$ to obtain “Efficiency” and “Speed” scores. In this way, a greater area inside the chart corresponds to a better-performing model.

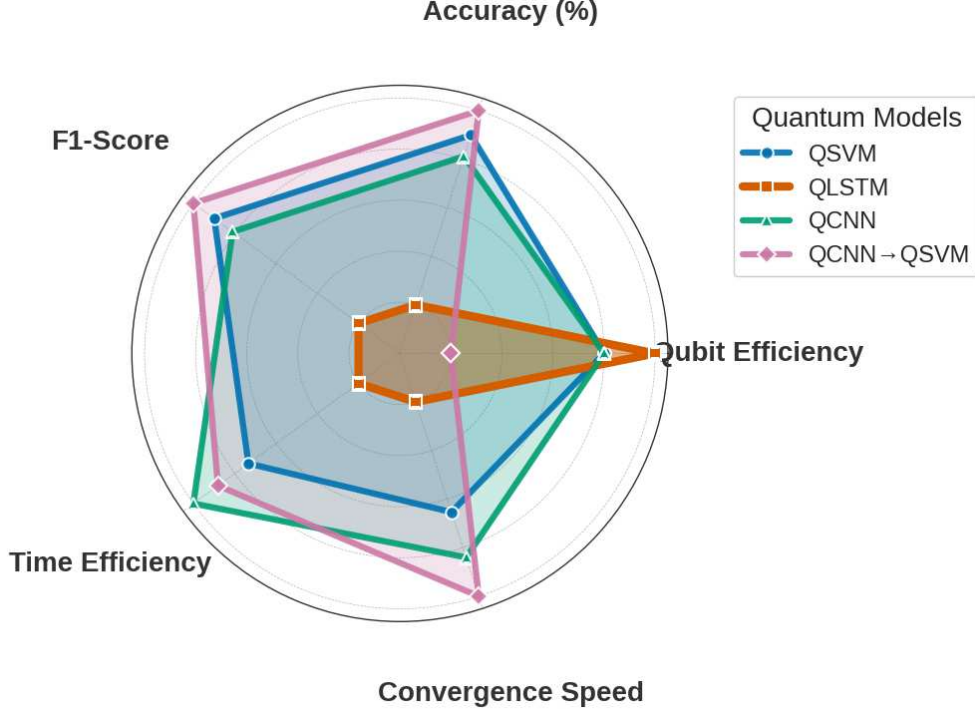


Fig. 8: Comparative Performance for All Models

The diagram clearly indicates that the “Hybrid QCNN→QSVM” has the best coverage, with a nearly perfect detection accuracy (99.5%) that converges to an optimum solution within only 22 epochs. This clearly validates that the efficiency advantage of the hybrid solution far surpasses the drawbacks of increased qubit usage.

In contrast, “QCNN” has a high “Time Efficiency” score (15.7s per epoch), which clearly indicates that the solution is highly reliable as a feature extractor. However, it is not optimal enough when compared to the standalone “QSVM” solution.

The profile of “QLSTM”’s shape clearly indicates that the solution’s limitation in qubit count (4) generated a drastic penalty on accuracy (40.0%).

5 Conclusion

Four quantum learning paradigms, namely QSVM, QCNN, QLSTM, and a hybrid QCNN→QSVM, were implemented. They were assessed for their efficacy in detecting deepfakes in audio signals. The findings show that the hybrid paradigm presents the optimal trade-off with a maximum accuracy of 99.5%. We implemented a hybrid of QCNN and QSVM. QCNN is great at finding patterns efficiently, while QSVM is very strong at making decisions. By putting them together, we addressed the problem of high costs usually attached to QSVM. We also improved how well the system sorts data compared to using QCNN alone. However, the QLSTM did not show an

efficient performance. This shows us that there are limitations. It shows that it is still difficult for it to take on complex tasks that deal with time and patterns. It should be noted that these findings were obtained within a noiseless simulation setup and a single-language task, and the effects of hardware decoherence and gate noise have yet to be considered. Future studies hope to include testing these architectures on NISQ devices and extending them to multi-language datasets. Their successful implementation within these realistic settings will be imperative for using quantum methods within classical networks. Hence, providing a firm foundation within quantum audio forensics.

Declarations

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Rahma Alam Samiha and Md. Tasbirul Alom. The first draft of the manuscript was written by Rahma Alam Samiha. Ajmery Sultana and Rubaiyat Islam contributed to the conceptualization, as well as the review and editing of the manuscript. All authors read and approved the final manuscript.

Data Availability

The datasets generated and/or analysed during the current study are available in the IEEE DataPort repository, <https://doi.org/10.21227/eyg4-v467>.

Statement of Financial Interests

No funds, grants, or other support was received to conduct this study.

Statement of Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

References

- [1] Kietzmann, J., Lee, L., McCarthy, I., Kietzmann, T.: Deepfakes: Trick or treat? *Business Horizons* **63**(2), 135–146 (2020)
- [2] Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. *ACM Computing Surveys* **54**(1), 1–41 (2021)
- [3] Yi, J., Zhang, X., Xu, B.: Voice deepfake detection: A survey. *IEEE Access* **10**, 24635–24649 (2022)
- [4] Samiha, R.A., Alom, M.T., Hossain, R., Islam, R., Sultana, A.: Leveraging deep learning methods for detecting deepfake speeches. In: 2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI), pp. 1–5 (2025). <https://doi.org/10.1109/ICMI65310.2025.11141035>

- [5] Zhang, C., Zhou, Y., Chen, Z.: Detection of audio deepfakes using spectral features. In: IEEE ICASSP, pp. 6359–6363 (2021)
- [6] Wang, X., Yamagishi, J., King, S.: A comparative study of speech deepfake detection methods. *Interspeech*, 3700–3704 (2020)
- [7] Al Badawy, E.A., Lyu, S.: Detecting ai-synthesized speech using deep neural networks. In: IEEE ICASSP, pp. 2487–2491 (2019)
- [8] Korshunov, P., Marcel, S.: Vulnerability assessment and detection of deepfake videos. *IEEE ICASSP*, 2527–2531 (2019)
- [9] Ma, X., Li, W., Li, H.: End-to-end detection of voice deepfakes using self-attention networks. *IEEE/ACM TASLP* **30**, 1748–1761 (2022)
- [10] Mollah, R., Rahman, M., Alam, R.: Quantum machine learning for bangla deepfake speech detection. Unpublished Manuscript (2023)
- [11] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. *Nature* **549**, 195–202 (2017)
- [12] Schuld, M., Killoran, N.: Quantum machine learning in feature hilbert spaces. *Physical Review Letters* **122**(4), 040504 (2019)
- [13] Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., Gambetta, J.M.: Supervised learning with quantum-enhanced feature spaces. *Nature* **567**(7747), 209–212 (2019)
- [14] Tacchino, F., Macchiavello, C., Gerace, D., Bajoni, D.: An artificial neuron implemented on an actual quantum processor. *npj Quantum Information* **5**(1), 26 (2019)
- [15] Liu, X., Xu, K., Li, J., Zeng, W.: Hybrid quantum-classical models for speech deepfake detection. *IEEE Transactions on Quantum Engineering* **3**, 1–12 (2022)
- [16] Ciliberto, C., Herbster, M., Pontil, M.: Quantum machine learning: a classical perspective. *Proceedings of the Royal Society A* **474**(2209), 20170551 (2020)
- [17] Rahman, R.M., Kaiser, M.S., Hossain, Z., Rahman, M.A., Al-Mamun, A., Yasmin, R.: Quantum ai: Harnessing the power of quantum computing for scalable and adaptive learning. *IEEE Access* **12**, 124230–124251 (2024) <https://doi.org/10.1109/ACCESS.2024.3467512>
- [18] Rahman, M.S., Hasan, M.M., Rahman, M.S., Khalil, I., Kaiser, M.S., Niyato, D.: Quantum machine learning for 6g network intelligence and adversarial threats. *IEEE Communications Standards Magazine* **9**(1), 12–23 (2025) <https://doi.org/10.1109/MCOMSTD.2025.3539955>

- [19] Alom, M.T., Samiha, R.A., Islam, R., Sultana, A.: Bengali Real and Deepfake Audio Dataset for Deepfake Detection. IEEE DataPort (2025). <https://doi.org/10.21227/eyg4-v467>
- [20] Pham, T.-L., *et al.*: Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. arXiv preprint (2024) <https://doi.org/10.48550/arxiv.2407.01777>
- [21] Kanwal, T., *et al.*: Fake speech detection using vggish with attention block. EURASIP Journal on Audio, Speech, and Music Processing **2024** (2024) <https://doi.org/10.1186/s13636-024-00348-4>
- [22] Yadav, A.K.S., *et al.*: Dsvae: Disentangled representation learning for synthetic speech detection. In: Proc. IEEE ICMLA (2023). <https://doi.org/10.1109/icmla58977.2023.00072>
- [23] Xie, Z., *et al.*: Fakesound: Deepfake general audio detection. arXiv preprint (2024) <https://doi.org/10.48550/arxiv.2406.08052>
- [24] Kreuk, G., *et al.*: Ai-synthesized voice detection using neural vocoder artifacts. arXiv preprint (2023) <https://doi.org/10.48550/arxiv.2304.13085>
- [25] Wolter, M.: Towards generalizing deep-audio fake detection networks. arXiv preprint (2023) <https://doi.org/10.48550/arXiv.2305.13033>
- [26] Pandey, A., Rudra, A.: Hybrid classical quantum learning model framework for detection of deepfake audio. Journal of Intelligent Systems (2025)
- [27] Lin, X., *et al.*: Quantum-trained convolutional neural network for deepfake audio detection. Quantum Machine Intelligence (2024)
- [28] Author, C., Author, D.: Quantum-powered deepfake detection. Future Internet (2025)
- [29] Chaharlang, J., Mosleh, M., Heikalabad, S.R.: Quantum reversible audio steganalysis using quantum schmidt decomposition and quantum support vector machine. Journal of Information Security and Applications **82**, 103755 (2024) <https://doi.org/10.1016/j.jisa.2024.103755>
- [30] Thakar, Y., Ghosh, B., Adeshra, V., Srivastava, K.: Performance analysis of hybrid quantum-classical convolutional neural networks for audio classification. In: Proc. IEEE ICCCNT (2024). <https://doi.org/10.1109/icccnt61001.2024.10725668>
- [31] Krishna, H., Sridevi, S., R., P.: Quantum kernel-aided support vector machine classifier for improved speech classification. In: Proc. IEEE ICCCNT (2023). <https://doi.org/10.1109/icccnt56998.2023.10307618>

- [32] Esposito, M., Uehara, G.S., Spanias, A.: Quantum machine learning for audio classification with applications to healthcare. In: Proc. IEEE IISA (2022). <https://doi.org/10.1109/IISA56318.2022.9904377>
- [33] Wu, Y., Adermann, E., Thapa, C., Camtepe, S., Suzuki, H., Usman, M.: Radio Signal Classification by Adversarially Robust Quantum Machine Learning (2023). <https://doi.org/10.21203/rs.3.rs-3842432/v1>
- [34] Author, X., Author, Y.: Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In: Proc. IEEE ICASSP (2021)
- [35] Author, Z., Author, W.: Design and analysis of a novel quantum-classical hybrid neural network for environmental sound classification. Applied Computing and Informatics (2024)
- [36] Author, Q., Author, R.: Parameterized quantum circuits for novel representation learning in speech emotion recognition. In: Proc. IEEE ICASSP (2025)
- [37] Katı, B.E., Küçüksille, E.U., Sarıman, G.: Enhancing deepfake detection through quantum transfer learning and class-attention vision transformer architecture. Applied Sciences **15** (2025) <https://doi.org/10.3390/app15020525>
- [38] Author, A., Author, B.: Framework for quantum-based deepfake video detection (without audio). Journal of Quantum Computing (2025)
- [39] Author, S., Author, T.: Hybrid quantum machine learning based human speech emotion recognition. Journal of Computational Acoustics (2025)