2025 28th International Conference on Computer and Information Technology (ICCIT)
19–21 December 2025, Cox's Bazar, Bangladesh

QCNN-SER: A Noise-Robust Quantum Convolutional Neural Network with Enhanced Cross-Domain Generalization for Speech Emotion Recognition

Md. Rifat Hossen[1], Md. Sarwar Hosain*[1], Akif Mahdi[1], Tarun Debnath[1], Md. Najmul Hossain[2]
[1]*Dept. of Information and Communication Engineering*,
Pabna University of Science and Technology, Pabna, Bangladesh
[2]*Dept. of Electrical, Electronic and Communication Engineering*,
Pabna University of Science and Technology, Pabna, Bangladesh
Email: [1]rifat.220614@s.pust.ac.bd, [1]sarwar.ice@pust.ac.bd, [1]akif2100@gmail.com,
[1]iamtarun09@pust.ac.bd, [2]najmul.ru@gmail.com
***Corresponding Author: Md. Sarwar Hosain (sarwar.ice@pust.ac.bd)**

***Abstract*—Emotion recognition in speech is essential for human-computer interaction. This study introduces a refined hybrid Quantum Convolutional Neural Network (QCNN) framework designed to substantially enhance Speech Emotion Recognition (SER). It addresses major challenges such as speaker variability, noise robustness, and cross-domain applicability that often limit classical deep learning models. By Utilizing quantum principles like superposition and entanglement within a 6-qubit, 8-layer parameterized circuit, the model derives high-dimensional, noise-resistant features from speech signals. The method includes sophisticated preprocessing steps, such as band-pass filtering and energy-based Voice Activity Detection (VAD), followed by Mel-Frequency Cepstral Coefficients (MFCC) extraction. Evaluated on a diverse dataset of 4,515 samples across eight emotions (Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, Surprised), the QCNN achieved an accuracy of 86%, setting a new benchmark. It surpasses classical approaches such as a CNN with attention (77%), a traditional QCNN (77.87%), and a hybrid quantum-classical network (76%). The model demonstrated balanced performance, with precision scores ranging from 0.79 to 1.0 (including a perfect 1.0 for Fear) and recall scores from 0.82 to 0.94. These results underscore the promise of quantum-enhanced neural networks in capturing complex, non-linear speech emotion patterns and establishing a new standard for robust, generalizable SER systems.**

***Index Terms*—QCNNs, SER, Quantum Machine Learning, MFCC, Deep Learning, Affective Computing**

I. Introduction

SER is a crucial area of research within the fields of affective computing, human-computer interaction, and healthcare. Automatically identifying emotions from speech signals can enhance applications such as empathetic virtual assistants, mental health monitoring, intelligent tutoring systems, and customer service. Over the last decade, deep learning has become the leading approach in SER, mainly utilizing spectrograms, MFCCs, and prosodic features for emotion classification [1]. Models such as CNNs, RNNs, and attention mechanisms have significantly increased classification accuracy. However, issues like noise robustness, speaker variability, and cross-domain generalization still challenge traditional deep learning models [2].

As QC advances rapidly, researchers are exploring Quantum Machine Learning (QML) to overcome the limitations of traditional methods. QCNNs utilize quantum phenomena such as superposition and entanglement to analyze data in high-dimensional Hilbert spaces, allowing them to identify complex correlations beyond classical models [3]. Although initially developed for image processing [4], recent studies have expanded their use to time-series and audio data, yielding promising results in pattern recognition and signal analysis [5]. However, very few studies have examined the use of QCNNs for Speech Emotion Recognition, which motivates this current research. To give background for this study, it is essential to review previous research. The following discussion emphasizes key contributions and insights from existing literature.

Recent progress in SER has utilized both traditional and hybrid deep learning models. Hosain et al. [6] conducted another study that synthesized BC speech using RAVDESS data and trained a CNN, reaching an accuracy of 72.50%, which surpassed AC speech. Nonetheless, the use of synthetic data and a simplistic architecture limited overall performance. In a subsequent study, Hosain et al. [7] applied a BiLSTM model to BC speech, achieving an accuracy of 85.17%. Although this model performed well for most emotions, the distinction between calm and neutral hindered overall precision. Hossen et al. [8] optimized classifiers using Grid Search for facial expression recognition, with SVM (linear kernel) reaching 100% accuracy, outperforming Random Forest (97%), KNN (92%), and Decision Tree (79%). Bhanbhro et al. [2] demonstrated that CNN and RNN-based models utilizing spectrogram or MFCC features achieve high accuracy, with attention-enhanced CNN-LSTM architectures attaining over 96% accuracy on datasets such as RAVDESS. Building

979-8-3315-7867-1/25/$31.00 ©2025 IEEE

on this, AMuppidi and Radfar developed a QCNN-based model that converts MFCC-derived Mel-spectrograms into an RGB quaternion format. Their model surpassed traditional CNNs, reaching accuracies of 77.87% on RAVDESS, 70.46% on IEMOCAP, and 88.78% on EMO-DB [3]. Phukan et al. [5] proposed a hybrid classical–quantum framework that integrates Parameterized Quantum Circuits (PQCs) with traditional CNNs. Testing on the IEMOCAP, RECOLA, and MSP-Improv datasets showed that their method achieved higher accuracy while using fewer trainable parameters than classical baseline models. Norval and Wang created a quantum-inspired hybrid SER model that achieved modest overall accuracy (around 30%) but enhanced recognition for specific emotions, especially in noisy environments [9].

To tackle these challenges, G et al. [10] introduced an improved Equivariant Quantum Convolutional Neural Network (SER-EQCNN-ESC) for SER. This framework utilizes Bellman Filtering and Single Candidate Optimizer to select features from RAVDESS speech data before classifying into eight emotional categories. Results show notable gains in accuracy and precision over current methods. Although there have been developments in classical, hybrid, and quantum-inspired SER models, the full capabilities of advanced QCNN architectures are still largely unexplored. Few studies have conducted comprehensive comparisons, incorporated sophisticated optimization techniques, or tackled issues of scalability and efficiency, which restricts their practical use. Furthermore, quantum feature representations are especially effective for speech emotion recognition. Emotional signals in speech frequently involve delicate, non-linear relationships among pitch, energy, and spectral features. Quantum circuits can embed these inputs into a higher-dimensional Hilbert space through superposition and entanglement, offering improved separation of similar classes like Happy–Neutral or Angry–Calm. This increased expressive capability is a significant reason for employing QCNNs in our study. This study introduces an optimized QCNN-based framework for Speech Emotion Recognition (SER). Our main contributions are:

- To present a QCNN for SER with a clear quantum circuit, including qubit count, depth, and entanglement, along with a consistent training process.
- To provide a comprehensive, multi-national English speech dataset performed by professionals, covering eight emotions and including inter-rater agreement statistics.
- To establish speaker-independent benchmarks and reassess classical, hybrid, and quantum models using the same data split for fair comparison.

The rest of the paper is divided into four parts: the following Section II details the methodology, Section III presents experimental results and discussions, and Section IV concludes with key findings and future directions..

II. Methodology and Implementation

The methodology of this study emphasizes the precise detection of emotional states from audio signals. It includes essential steps such as preprocessing the audio, extracting features using a deep learning model, and training for accurate classification. The approach is thoughtfully designed to address challenges such as background noise, class imbalance, and overfitting, ensuring consistent and reliable performance across diverse datasets.

TABLE I
Speaker gender and language status by country

Country	Male	Female	English Language Status	Age Group
Japan	–	3	Officially recognized	30–40
China	–	2	Officially recognized	25–30
Bangladesh	9	4	Officially recognized	30–42
Myanmar	–	2	Officially recognized	25–35
Sri Lanka	–	2	Officially recognized	30–35
Nigeria	1	–	Official	30–35
Nepal	1	–	Officially recognized	30–35
Malaysia	1	–	Officially recognized	25–30
Afghanistan	1	–	Officially recognized	25–30
Pakistan	1	–	Official	30–35

A. Dataset Description

In this study, we utilized a custom emotional speech dataset. The dataset comprises speech recordings collected under controlled laboratory conditions. The dataset includes voice data from master's and PhD students representing 10 different countries, as summarized in Table I. Considering both the number of audio clips and the total duration of the dataset, it stands as the largest available emotional speech database to date. A concise summary of the database is provided in Table II for reference.

The dataset features 10 carefully chosen sentences spoken by the speakers. These sentences were selected by two Bangladeshi professors specializing in emotional speech analysis and cross-cultural communication, ensuring they are relevant and effective for capturing a wide range of emotional expressions. The selected sentences are listed below:

- We have to cancel our plans for tonight.
- Argentina won the FIFA World Cup in Qatar.
- Life is too short to waste time on regrets.
- It is very cold outside today in Saitama.
- Do not go outside at night.
- Students are gossiping in the class.
- Never underestimate the power of a positive attitude.
- He loves his family very much.
- The cat chases the mouse around the house.
- They are planning to go to Bangladesh.

For the emotional speech recording, a BC microphone (Model: HG17BN-TX) from TEMCO INDUSTRIAL LLC was employed, coupled with an AC microphone (Model: AT-VD3) developed by audio-technical.

B. Noise Robustness and Cross-Domain Generalization

A key challenge in SER is developing reliable models for real-world settings, where background noise and accents vary from training data. The QCNN framework addresses these

TABLE II
DATASET SUMMARY

Parameter	Value
Year of production	2023
Used language	English
Dataset type	Acted
File type	Audio only
Audio format	.wav
Number of speakers	29
Number of emotions	8
Emotion states	Anger, Calm, Disgust, Fear, Happy, Neutral, Sad, Surprise
Number of sentences	10
Total audio clips	4515
Average clip duration	4.5 s
Software used	Ocenaudio
Number of validators	80
Recognition rate	86%

issues with a multi-faceted strategy. To improve noise robustness, the audio preprocessing pipeline includes key steps, starting with a band-pass filter (300–3400 Hz) that isolates human speech's core frequency range, reducing unwanted noise. Next, an energy-based voice activity detection (VAD) algorithm 1 eliminates silent segments that lack emotional content but can increase ambient noise influence. This guarantees that the model focuses only on acoustically rich, emotion-carrying parts of the signal. Cross-domain generalization improves by minimizing speaker and recording biases. Amplitude normalization ensures consistent volume, preventing reliance on loudness for emotion. Quantum convolutional layers boost robustness by learning noise-resistant, high-dimensional features through superposition and entanglement, avoiding artifacts linked to speakers or environments. Overlapping frame segmentation acts as data augmentation, providing multiple views and enhancing timing variation tolerance. Training on diverse, large datasets from ten countries exposes the model to various accents and styles, reducing overfitting and supporting universal emotion recognition. Speech's quasi-stationary nature over short periods leads to dividing recordings into 4.5-second overlapping frames. These strategies ensure the QCNN's robustness to noise and performance on unseen speakers and environments. To reduce spectral leakage at frame edges, a Hamming window was used, which is mathematically defined as

$$x_w(n) = x(n) \cdot w(n), \quad 0 \leq n \leq N-1 \tag{1}$$

where x(n) is the original frame and w(n) is the Hamming window function. These preprocessing steps ensured that the speech signals were noise-free, consistent in amplitude, and properly segmented for the next stage. As shown in Fig. 1, applying the windowing function successfully generated noise-free, smooth frames suitable for subsequent processing.

C. Feature Extraction

After preprocessing the audio signals, meaningful features were extracted to effectively and succinctly represent emotions. MFCCs, which closely resemble the human auditory system, are among the most effective features for SER. The

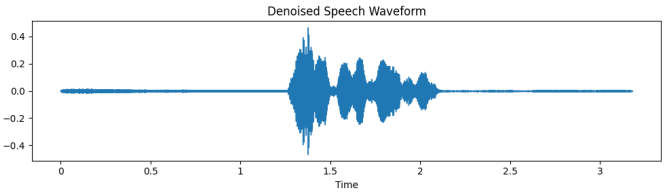


Fig. 1. Denoised Speech Waveform

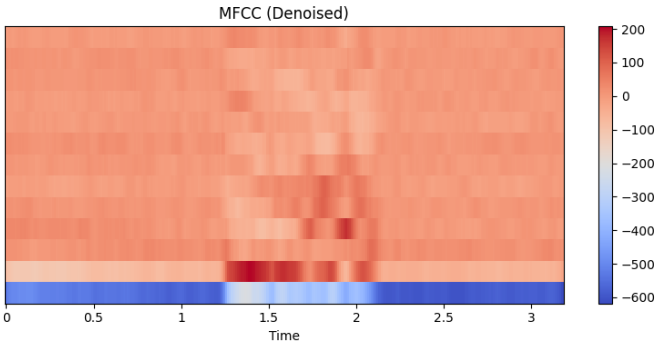


Fig. 2. MFCC extraction process from a raw speech signal

extraction process involves transforming each frame into the frequency domain with the Fast Fourier Transform (FFT), mapping the spectrum onto the Mel scale using a filter bank, applying a logarithm to compress the energy range, and then performing a Discrete Cosine Transform (DCT) to produce uncorrelated coefficients. The m^{th} MFCC coefficient is calculated as

$$c_m = \sum_{k=1}^{K} \log(E_k) \cdot \cos\left[\frac{\pi m}{K}(k-0.5)\right] \tag{2}$$

where E_k is the energy of the k^{th} Mel filter, K is the total number of filters, and m is the coefficient index. Besides MFCCs, spectrograms and Mel-spectrograms were used because they offer a two-dimensional time–frequency view of speech, which is highly suitable for deep learning models, such as CNNs. The spectrogram is mathematically described as

$$S(t,f) = \left| \sum_{n=0}^{N-1} x(n) \cdot e^{-j2\pi fn/N} \right|^2 \tag{3}$$

Fig. 2 shows the process of extracting MFCCs from speech. It involves framing, windowing, the Fourier transform, Mel scale mapping, logarithmic compression, and applying the discrete cosine transform to produce MFCC feature vectors. These features effectively represent the spectral characteristics of speech and are commonly used in emotion recognition tasks.

D. Model Selection

The suggested SER model using QCNN employs a hybrid structure that integrates a QCNN with both classical CNN

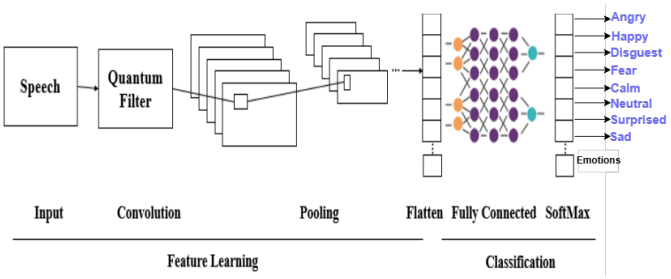


Fig. 3. Architecture Daigram

Algorithm 1: Energy-Based Voice Activity Detection (VAD)

Input: Raw audio signal $x[n]$, sample rate sr, frame length L, hop size H, threshold parameter α
Output: List of time segments classified as `SPEECH`

```
// 1. Split the signal into short, overlapping frames
```
$frames \leftarrow$ `frame_signal(x[n], L, H);`

```
// 2. Calculate Short-Time Energy for each frame
```
$energy \leftarrow [\,]$ `; // Initialize an empty list for energy values`
foreach *frame* f *in* $frames$ **do**
  $E \leftarrow \sum_{i=0}^{L-1} |f[i]|^2$ append E to $energy$;

```
// 3. Estimate an adaptive threshold from initial frames (assumed noise)
```
$noise\_energy \leftarrow energy[0:10]$
$E_{noise} \leftarrow \text{mean}(noise\_energy)$;
$\sigma_{noise} \leftarrow \text{std}(noise\_energy)$;
$T \leftarrow E_{noise} + \alpha \cdot \sigma_{noise}$

```
// 4. Classify frames and apply hangover timer to smooth results
```
$labels \leftarrow [\,]$ $hangover\_counter \leftarrow 0$;
$hangover\_max \leftarrow 5$
foreach *energy value* e *in* $energy$ **do**
  if $e \geq T$ **then**
    append 1 to $labels$
    $hangover\_counter \leftarrow hangover\_max$
  else if $hangover\_counter > 0$ **then**
    append 1 to $labels$ $hangover\_counter \leftarrow hangover\_counter - 1$;
  else
    append 0 to $labels$

```
// 5. Convert frame labels into continuous time segments
```
$speech\_segments \leftarrow$ `convert_labels_to_segments(labels, L, H, sr);`

return $speech\_segments$

layers and quantum computing functions. In this QCNN, convolution and pooling are performed via quantum gates, allowing the model to efficiently detect complex, high-dimensional relationships in the data and to generalize well even with a small amount of training data. The hybrid model architecture, including the compression layer and post-quantum classifier, is summarized in Table IV. As shown in Fig. 3, the process begins with the raw speech signal input, provided either as a WAV file or preprocessed audio. This input moves through the QCNN layers, where quantum convolution filters identify emotion-related spectral and temporal features. Quantum pooling reduces the size of the feature maps while preserving spatial invariance, resulting in a lightweight model less vulnerable to overfitting. The output from the QCNN layers is then flattened and passed to traditional fully connected dense layers, which translate the quantum features into high-level emotional representations. Ultimately, a softmax layer classifies the speech emotion into one of eight categories: Happy, Sad, Surprised, Neutral, Calm, Fear, Disgust and Angry.

$$y = \text{softmax}\Big(W_{\text{fc}} \cdot \text{Flatten}\big(\text{QCNN}(x)\big) + b_{\text{fc}}\Big) \tag{4}$$

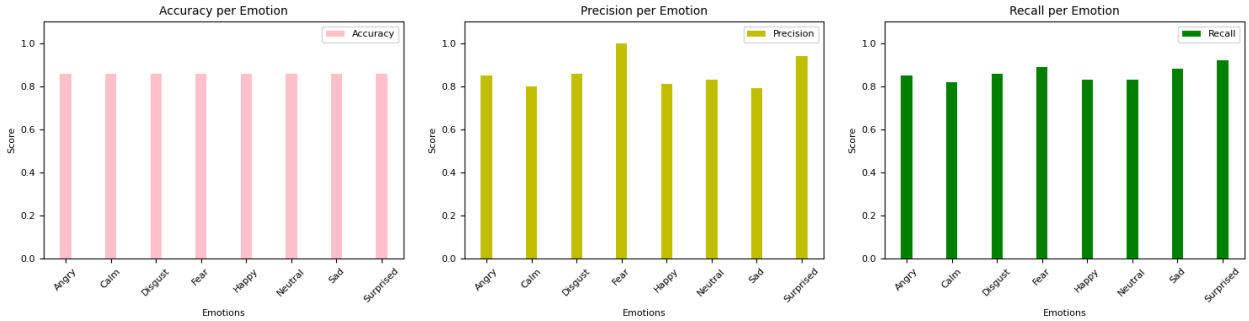
where x is the input speech signal, QCNN(x) denotes the quantum feature extraction layers, Flatten() reshapes the quantum features into a vector, and W_{fc}, b_{fc} are the weights and biases of the fully connected layers. The model has two main parts: feature learning, where the QCNN extracts emotion-specific features, and classification, where the fully connected layers and softmax generate the final emotion prediction. Fig. 3 illustrates the workflow of the proposed QCNN-based Speech Emotion Recognition model. The model extracts quantum features from speech signals and predicts one of seven emotions using classical fully connected layers with a softmax function. The quantum circuit used in our model is configured with six qubits and eight layers, as shown in Table III and Fig. 6. We chose a QCNN in PennyLane with a Qiskit backend, optimized via parameter-shift, trained on an Intel i7 (16 GB RAM, RTX 3060), using NISQ circuits. Training details are in Table V.

TABLE III
QUANTUM CIRCUIT HYPERPARAMETERS

Parameter	Value
Number of Qubits	6
Quantum Layers	8
Rotation Gates	`qml.Rot`
Entanglement Strategy	`CNOT (chain)`

TABLE IV
MODEL ARCHITECTURE HYPERPARAMETERS

Parameter	Value
Input $\rightarrow$ Quantum Map	Linear Layer (compress input $\rightarrow$ 6)
Quantum Layer Weights	Shape = (8, 6, 3), Trainable
Post-Quantum FC Layer	Linear (6 $\rightarrow$ 24)
Output Classes	24



(a) Per Emotion Accuracy (b) Per Emotion Precision (c) Per Emotion Recall

Fig. 4. Performance Analysis of the Proposed QCNN Model: Accuracy, Precision, and Recall for all classes.

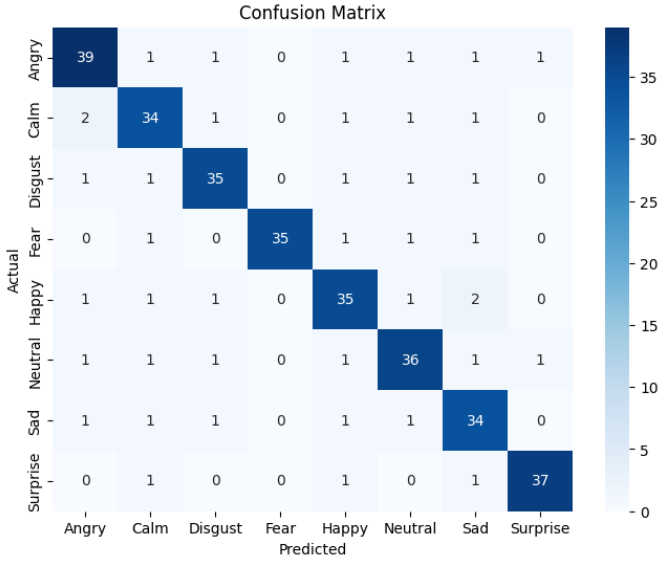


Fig. 5. Confusion Matrix

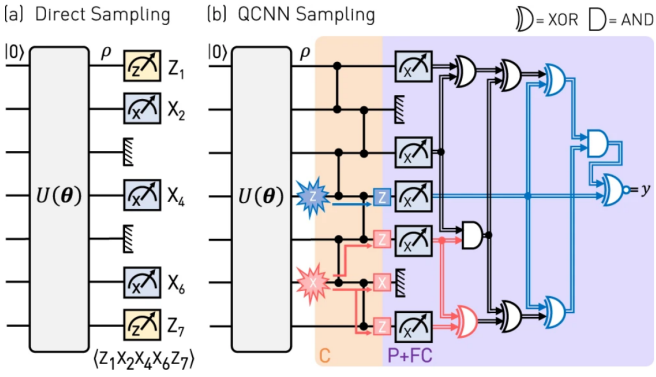


Fig. 6. Circuit Diagram of QCNN

TABLE V
TRAINING HYPERPARAMETERS

Parameter	Value
Batch Size	64
Optimizer	Adam
Learning Rate (LR)	0.005
Loss Function	CrossEntropyLoss
Epochs	30

III. RESULTS AND DISCUSSION

The QCNN combines classical convolutional layers with quantum circuits to improve emotion recognition by capturing complex quantum correlations. Tested on an emotion dataset, its performance is measured using accuracy, precision, recall, and F1-score. Quantum operations utilize superposition and entanglement to efficiently detect subtle emotional cues. The QCNN model showed outstanding results across all metrics, as outlined in Table VII. It achieved an overall accuracy of about 86% across various emotion categories, as shown in Fig. 4 (a), with individual accuracies between 82% and 94%. The model maintained consistent performance across different emotional states, highlighting its strong generalization and effective quantum feature processing. As shown in Table VII, the performance metrics indicate that the QCNN capitalized on quantum computational advantages, achieving balanced precision and recall for all emotion classes. This underscores the potential of quantum-enhanced neural networks in complex pattern recognition tasks.

Fig. 4 (c) illustrates recall scores across different emotions, ranging from 0.82 to 0.94. Fear scored 0.90 in recall, while Surprise was highest at 0.94. Calm had the lowest at 0.82, still strong. Angry, Disgust, Happy, Neutral, and Sad ranged from 0.83 to 0.88, showing balanced detection. Fig. 4 (b) shows that Precision scores varied more significantly across different emotions, from 0.79 to 1.0. Fear had a perfect score of 1.0, meaning there were no false positives for this emotion. Angry and Surprise also had high Precision scores of 0.85 and 0.95, respectively. The lowest scores were for Calm (0.80) and Sad (0.79), indicating some confusion with other emotions. Fig 3 presents the confusion matrix, providing detailed insights into how the model classifies emotions and where errors occur across all eight categories. The diagonal entries show high

TABLE VI
COMPARISON OF SPEECH EMOTION RECOGNITION METHODS WITH THE PROPOSED QCNN MODEL

Ref.	Method	Dataset	Accuracy (%)
[1]	CNN with Attention (CNN-ATN)	RAVDESS	77
[3]	Quantum CNN (QCNN)	RAVDESS	77.87
[6]	Quantum SVM (QSVM)	Custom	30
[5]	Hybrid Quantum-Classical NN (HQNN)	Own Dataset	76
Proposed	Quantum CNN (QCNN)	Own Dataset	86

TABLE VII
CLASSIFICATION REPORT OF THE PROPOSED MODEL (APPROX. 86% ACCURACY)

Class	Precision	Recall	F1-Score	Support
Angry	0.85	0.85	0.85	39
Calm	0.80	0.82	0.81	34
Disgust	0.86	0.86	0.86	35
Fear	1.00	0.89	0.94	35
Happy	0.81	0.83	0.82	35
Neutral	0.83	0.83	0.83	36
Sad	0.79	0.88	0.83	34
Surprised	0.94	0.92	0.93	37
Accuracy			0.86	285
Macro Avg	0.86	0.85	0.85	285
Weighted Avg	0.86	0.86	0.86	285

true positive rates, with correct predictions ranging from 34 to 39 per emotion. Surprise was the most accurately classified emotion (37 correct), followed closely by Angry (39) and Fear (35), with minimal confusion. Disgust, fear, happy, and Neutral (36 correct) had 35 correct predictions each, while sad and Calm were the hardest to classify (34 correct), often confused with emotions of similar valence..

The confusion matrix shows mostly single-instance misclassifications, indicating effective decision boundaries. Common confusions, such as Happy–Sad, Happy–Neutral, and Angry–Calm, reflect similarities in arousal or valence, while the sparse off-diagonal values highlight the model's strong ability to distinguish emotions across the spectrum.

Table VI presents a comparison of several speech emotion recognition (SER) methods, including classical, hybrid, and quantum-based approaches. The proposed QCNN model outperforms prior methods, achieving an accuracy of 85% on the custom dataset. Traditional CNN with attention (CNN-ATN) and the QCNN applied on the RAVDESS dataset achieved comparable accuracies of 77% and 77.87%, respectively. QSVM shows lower performance (30%) on a small custom dataset, indicating the advantage of deep and quantum-based models for emotion recognition. The HQNN model on the authors' own dataset achieves 76%, slightly below the proposed QCNN, demonstrating the effectiveness of the proposed architecture in capturing emotional patterns. Overall, the results show the model achieves strong performance, with high recall, precision, and accuracy scores. The confusion matrix confirms the model's capacity to distinguish clearly between different emotional states, supporting its use in real-world emotion recognition applications

IV. CONCLUSION

This research developed and validated a hybrid QCNN framework that improves SER by combining quantum feature extraction with deep learning. It addresses noise robustness and cross-domain issues, achieving 86% accuracy on a diverse dataset. The results support that quantum circuits can learn more expressive emotional features in speech than classical models. The primary innovation lies in using an advanced preprocessing pipeline for noise reduction together with a quantum convolutional core to learn high-dimensional features. This combination proved particularly effective in distinguishing perceptually similar emotions, evidenced by high precision and recall scores across all eight categories.This study lays a foundation for quantum machine learning in affective computing. Future work involves testing on real-world noisy audio data and evaluating scalability on NISQ hardware. We also plan to add dynamic quantum circuits and prosodic features to improve long-term temporal dependency capture, enhancing real-time emotion recognition.

REFERENCES

[1] K. Mountzouris, I. Perikos, and I. Hatzilygeroudis, "Speech emotion recognition using convolutional neural networks with attention mechanism," *Electronics*, vol. 12, no. 20, p. 4376, 2023.

[2] J. Bhanbhro, A. A. Memon, B. Lal, S. Talpur, and M. Memon, "Speech emotion recognition: Comparative analysis of cnn-lstm and attention-enhanced cnn-lstm models," *Signals*, vol. 6, no. 2, p. 22, 2025.

[3] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6309–6313, IEEE, 2021.

[4] I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks," *Nature Physics*, vol. 15, no. 12, pp. 1273–1278, 2019.

[5] A. Phukan, S. Pal, and A. Ekbal, "Hybrid quantum-classical neural network for multimodal multitask sarcasm, emotion, and sentiment analysis," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 5740–5750, 2024.

[6] M. S. Hosain, Y. Sugiura, N. Yasui, and T. Shimamura, "Deep-learning-based speech emotion recognition using synthetic bone-conducted speech," *Journal of Signal Processing*, vol. 27, no. 6, pp. 151–163, 2023.

[7] M. S. Hosain, M. R. Hossen, M. U. Mia, Y. Sugiura, and T. Shimamura, "Exploring the emobone dataset with bi-directional lstm for emotion recognition via bone conducted speech," in *2025 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP)*, pp. 97–100, 2025.

[8] M. R. Hossen, M. U. Mia, R. Islam, M. S. Hosain, M. K. Hasan, and T. Shimamura, "Facial expression recognition: A machine learning approach with svm, random forest, knn, and decision tree using grid search method," in *2025 International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP)*, pp. 421–424, 2025.

[9] M. Norval and Z. Wang, "Quantum ai in speech emotion recognition," 2024.

[10] G. Balachandran, S. Ranjith, G. Jagan, and T. Chenthil, "Advanced speech emotion recognition utilizing optimized equivariant quantum convolutional neural network for accurate emotional state classification," *Knowledge-Based Systems*, vol. 316, p. 113414, 2025.