Journal of Advanced College of Engineering and Management

# Nepali Speech Emotion Recognition Using Variational Quantum Circuits

[1]Nishchal Pokhrel, [2*]Nanda Bikram Adhikari
[1,2]*Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, Tribhuvan University, Nepal*
*Corresponding email: [*]adhikari@ioe.edu.np*

*DOI: 10.3126/jacem.v12i01.93927*

***Abstract***
Speech emotion recognition (SER) is an active area of research, yet existing work has focused almost exclusively on high-resource languages, leaving Nepali — with over 32 million first- and second-language speakers worldwide — without any published SER study or emotional-speech corpus. This paper addresses that gap along two dimensions. First, we construct a Nepali emotional-speech dataset comprising 600 utterances across three emotion classes (happy, sad, neutral), validated by 117 native listeners whose mean recognition accuracy is 91.5%. Second, on this corpus, we evaluate a fully quantum data-reuploading variational quantum circuit (VQC) classifier with trainable SU(2) encoding on nine qubits, and compare it directly against two classical baselines — a random forest and a multilayer perceptron — on the same PCA(27) feature pipeline and stratified 480/120 split. A staged hyperparameter search covering circuit depth, learning rate, optimizer, and batch size identifies an optimal VQC configuration of ten layers and 540 trainable parameters, which attains 90.83% test accuracy and a macro-F1 of 0.908. Gradient-norm analysis confirms the absence of barren plateaus during training. Both classical baselines outperform the VQC under this protocol (Random Forest 95.00%, MLP 99.17%); however, a leave-one-speaker-out robustness check shows that classical accuracy collapses by approximately one-third under this evaluation, indicating that a substantial portion of the classical advantage reflects speaker-level information leakage.

***Keywords****—Data Reuploading Classifier, Nepali Emotion Dataset, SER, VQC*

## 1. INTRODUCTION

Speech emotion recognition (SER) is the computational task of automatically identifying a speaker's affective state from acoustic properties of the voice—including pitch contour, energy, rhythm, spectral shape, and voice quality—which carries emotional information independent of lexical content [1]. As a core component of affective computing, SER has attracted substantial research interest across diverse application domains, including clinical mental health screening, customer service, adaptive educational technology, and driver safety monitoring [2]. Despite these commercial and scientific interests, the progress of SER technology remains almost entirely confined to a small number of

high-resource languages—predominantly English, German, and Mandarin—leaving billions of speakers of other languages without access to emotion-aware AI systems.

This linguistic imbalance is particularly great for South Asian languages. A comprehensive review of SER progress across Indo-Aryan and Dravidian language families confirmed that many major languages in the region remain entirely unaddressed [3]. Nepali—an Indo-Aryan language spoken by over 32 million first-language speakers and approximately 40 million worldwide, serving as the official language of Nepal and recognised in several Indian states—has no published peer-reviewed SER study and no validated emotional speech dataset. Cross-lingual transfer approaches are inadequate because prosodic patterns, phonetic inventories, and cultural norms for emotional expression differ markedly across language families; models trained on English corpora cannot reliably generalise to Nepali speech. Even where language-specific datasets exist, conventional deep learning architectures—convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and Transformers—employ hundreds of thousands to millions of trainable parameters and frequently overfit when applied to datasets containing only hundreds to a few thousand utterances, the scale typical of newly constructed low-resource corpora [4].

Variational quantum circuits (VQCs) offer a theoretically compelling, parameter-efficient alternative. A VQC operates within a $2^n$-dimensional Hilbert space for n qubits yet requires only O(n) trainable parameters; quantum-mechanical properties such as superposition, entanglement, and interference enable complex feature interactions that would demand exponentially more classical parameters to represent [5]. Caro et al. [6] proved that the generalisation error of a quantum model with T trainable gates scales as $\sqrt{(T/N)}$, and demonstrated that a quantum convolutional network achieved over 97% accuracy with only 80 training samples—establishing that fewer parameters can yield better generalisation from limited data, a property directly relevant to low-resource SER. Despite these advantages, no study has applied a standalone pure VQC as an SER classifier, and no SER research of any kind exists for Nepali or any other South Asian language.

The main contributions of this paper are as follows: (1) we introduce the first human-validated Nepali emotional-speech dataset, comprising 600 recordings (10 speakers × 10 sentences × 3 emotions × 2 takes) with perceptual validation from 117 native listeners at a mean recognition accuracy of 91.5%; and (2) we present the empirical evaluation of a pure data-reuploading VQC classifier for Nepali speech emotion recognition, including a staged hyperparameter grid search over depth, learning rate, optimizer, and batch size, and a direct classical comparison against random forest and multilayer-perceptron baselines on the same feature pipeline and split.

## 2. RELATED STUDIES

### A. Classical and Deep Learning Approaches to SER

The conventional SER pipeline extracts hand-crafted acoustic features—Mel-frequency cepstral coefficients (MFCCs), mel-scale spectrograms, chroma vectors, and spectral contrast—and feeds them into classifiers ranging from support vector machines and random forests to CNNs, LSTMs, and Transformer models [2]. Mashhadi and Osei-Bonsu [4] compared a one-dimensional CNN against a random forest on a combined corpus of RAVDESS, CREMA-D, TESS, and SAVEE, and reported that the random forest with recursive feature elimination achieved 69% accuracy versus only 61% for the CNN. The authors explicitly attributed the CNN's inferior performance to overfitting caused by limited dataset size relative to model complexity—a finding that directly motivates the exploration of low-parameter alternatives for small corpora. More recently, self-supervised pre-trained models have established new benchmarks: fine-tuned HuBERT achieves 79.58% weighted accuracy on IEMOCAP in speaker-dependent settings [7], while the emotion2vec framework demonstrated competitive performance across ten languages using only linear downstream layers. However, these advances rely almost exclusively on large English, German (EMO-DB [8]), and Mandarin datasets, reinforcing a persistent linguistic imbalance in the field.

### B. Low-Resource and South Asian Language SER

A systematic review of SER progress across Indo-Aryan and Dravidian language families found that while Hindi, Bengali, and Urdu have received limited research attention, many major South Asian languages remain entirely unaddressed [3]. For Nepali specifically, natural language processing research has produced the NepBERT, a language model and crowd-sourced speech corpora for automatic speech recognition, but these resources target text and speech-to-text tasks rather than emotional speech analysis. Cross-lingual SER studies have consistently demonstrated that models trained on one language transfer poorly to others due to differences in prosodic patterns, phonetic inventories, and culturally determined norms for emotional expression—underscoring the fundamental necessity of language-specific datasets and models rather than cross-lingual transfer from existing corpora.

### C. Variational Quantum Circuits for Classification

Quantum machine learning, and specifically VQCs, has emerged as a theoretically grounded framework for parameter-efficient classification. Cerezo et al. [5] provided a comprehensive review of variational quantum algorithms, establishing VQCs as trainable models that leverage superposition and entanglement to explore high-dimensional feature spaces with far fewer parameters than equivalent classical networks. The expressivity of VQCs is critically governed by the data-encoding strategy: Schuld et al. [9] proved that quantum classifiers are natural partial Fourier

series in the input data, with the encoding gates determining the accessible frequency spectrum. Pérez-Salinas et al. [10] further demonstrated that data re-uploading—the technique of repeatedly embedding input data at multiple circuit layers—enables even a single qubit to serve as a universal classifier. Caro et al. [6] provided the first rigorous generalisation bound for quantum models, showing that the generalisation error scales as $\sqrt{(T/N)}$ where T is the number of trainable gates and N is the training set size, and empirically demonstrated that quantum convolutional networks can achieve over 97% accuracy from only 80 training samples. Rath and Date [11] further showed through systematic comparison of classical-to-quantum mapping techniques that encoding strategy has a decisive impact on downstream classification accuracy, making it a critical design variable for any VQC-based system.

### D. Quantum Approaches to Speech Emotion Recognition

Quantum approaches to SER remain in their early stages. Rajapakshe et al. [12], in the most comprehensive work to date, integrated parameterised quantum circuits into a classical CNN pipeline for SER on IEMOCAP, RECOLA, and MSP-IMPROV, achieving a 50.34% reduction in trainable parameters while improving classification accuracy. Critically, however, their architecture is a hybrid quantum-classical system in which quantum circuits serve as intermediate representation layers within a conventional neural network, not as a standalone classifier. Norval and Wang [13] applied VQC, QSVM, and QAOA-based classifiers to a custom Afrikaans emotional speech corpus, reporting 41–43% test accuracy under ideal simulation—significantly below their CNN-LSTM baseline of 73.9%. Additional hybrid approaches include quantum federated learning for vehicular SER and quantum-enhanced echo state networks, but in each case quantum components remain embedded within larger classical architectures. A critical gap thus persists: no study has evaluated a pure VQC—without any classical neural network layers—as a standalone SER classifier, and the systematic impact of quantum encoding strategies on audio emotion classification has not been investigated despite encoding being a theoretically decisive factor [9], [11].

## 3. METHODOLOGY

The overall pipeline proceeded in five stages: dataset construction and perceptual validation, audio preprocessing and feature extraction, dimensionality reduction and encoding preparation, variational quantum circuit design, and training with systematic hyperparameter evaluation. A generalized block diagram of the methodology is presented in Figure 1.

### A. Dataset Construction and Perceptual Validation

Since no publicly available emotional speech corpus exists for the Nepali language, a purpose-built dataset was constructed for this study. The Nepali-SER dataset comprised

600 audio samples recorded by 10 male native Nepali speakers, spanning three target emotion classes—Happy, Sad, and Neutral—with a perfectly balanced class distribution of 200 samples per emotion.

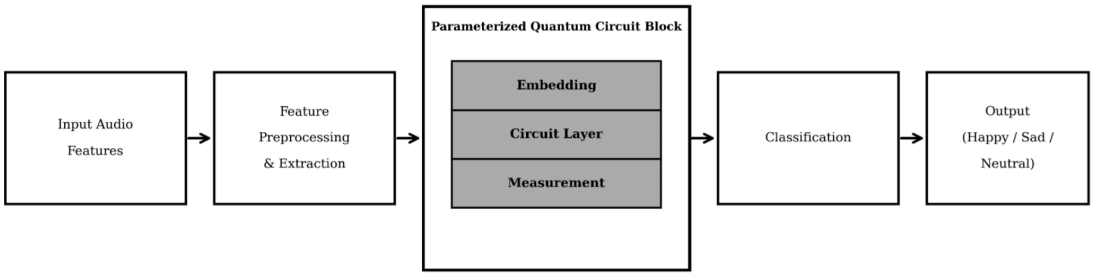


**Figure 1**: Block diagram of the proposed methodology

**a. Recording Protocol and Dataset Composition**

Each speaker recorded 10 semantically neutral sentences—selected to carry no inherent emotional bias—expressing each of the three target emotions with 2 takes per emotion per sentence, yielding 60 samples per speaker. All recordings were captured using Audacity 3.x with standardised settings of 44.1 kHz sample rate, 16-bit PCM encoding, mono channel, and WAV format. Speakers were positioned 15–20 cm from the microphone in a quiet indoor environment, with input levels calibrated to −12 to −6 dB to prevent clipping while ensuring adequate signal strength.

**Table 1**: Semantically neutral Nepali sentences used for emotional speech recording

| S.N. | Sentence (Nepali) |
|---|---|
| 1 | म हरेक बिहान सात बजे उठ्छु |
| 2 | मेरो घर यहाँबाट पाँच किलोमिटर टाढा छ |
| 3 | उसले आज बजारमा नयाँ कपडा किन्यो |
| 4 | म भोलि पोखरा जाने योजना बनाउँदैछु |
| 5 | हाम्रो विद्यालयमा धेरै विद्यार्थीहरू पढ्छन् |
| 6 | उनीहरूले गत हप्ता नयाँ फोन किनेका छन् |
| 7 | म प्रत्येक दिन चिया र रोटी खान्छु |
| 8 | यो पुस्तक पढ्न धेरै रोचक छ |
| 9 | मलाई संगीत सुन्न धेरै मन पर्छ |
| 10 | उनको काम साँच्चै रामो भएको छ |

Each recording underwent two-stage quality control: **technical verification** (absence of clipping, consistent volume, and absence of background noise) and **content verification** (correct pronunciation and appropriate emotional expression). Samples failing either criterion were re-recorded.

### b. File Naming Convention

Files were named according to the structured convention *A{XX}_S{YY}_{E}_T{Z}.wav*, encoding four metadata fields: actor identity, sentence number, emotion class, and take number. Table 2 presents the complete specification. As a representative example, the filename *A03_S07_H_T2.wav* identifies Actor 3, Sentence 7, emotion class Happy, second take. The full dataset of 600 files is computed as: 10 actors × 10 sentences × 3 emotions × 2 takes = 600.

**Table 2**: File naming convention for the Nepali-SER dataset

| Field | Token | Values | Description | Example |
|---|---|---|---|---|
| Actor ID | A{XX} | 01–10 | Zero-padded speaker index (10 speakers) | A03 |
| Sentence ID | S{YY} | 01–10 | Zero-padded sentence index (10 sentences) | S07 |
| Emotion class | {E} | H, S, N | H = Happy, S = Sad, N = Neutral | H |
| Take number | T{Z} | 1, 2 | Recording attempt number (2 takes per utterance) | T2 |
| File format | .wav | WAV | 16-bit PCM, 44.1 kHz, mono channel | .wav |

### c. Dataset Distribution

The dataset was designed with a perfectly balanced class distribution to eliminate label-frequency bias from the classification task, with each of the three emotion classes containing exactly 200 utterances (33.3% of total). Figure 2 presents representative amplitude waveforms for each emotion class, illustrating the characteristic acoustic differences: Happy speech exhibits consistently higher energy and greater amplitude variation; Sad speech shows lower intensity with a compressed dynamic range; and Neutral speech displays comparatively flat prosodic characteristics.

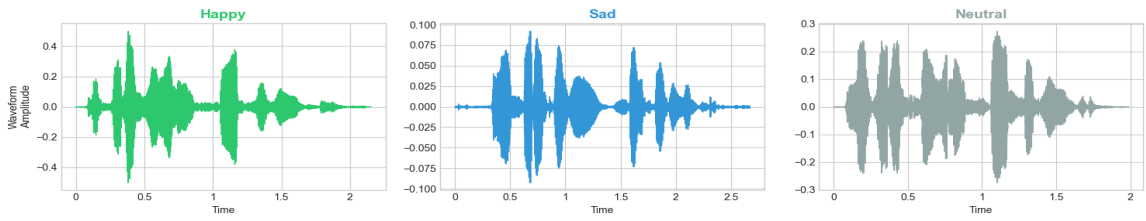


**Figure 2**: Representative amplitude waveforms for each emotion class.

### d. Perceptual Validation

To establish the emotional validity of the dataset, a perceptual validation study was conducted with 117 independent native Nepali listeners, generating 1,170 individual emotion judgments across 50 randomly selected clips. The evaluation forms were distributed digitally via five online questionnaires hosted on GitHub Pages. Respondents represented geographic diversity spanning six of seven provinces of Nepal and eight distinct mother tongues, with 76% identifying Nepali as their primary language.

The study achieved an overall human recognition accuracy of 91.5%, with per-class accuracies of 96.3% (Neutral), 90.6% (Happy), and 88.0% (Sad), as summarised in Table 3. The most frequent misclassification was Sad perceived as Neutral (10.0%), consistent with the known acoustic overlap between low-arousal emotional states, where reduced energy and flattened prosody create perceptual ambiguity at the Sad–Neutral boundary. All other confusion rates remained below 7%. The overall accuracy of 91.5% established a human performance upper bound against which computational models were subsequently evaluated.

**Table 3**: Perceptual validation results — per-class human recognition accuracy and primary confusion pattern.

| Emotion class | Correctly recognised (%) | Primary confusion | Confusion rate (%) |
|---|---|---|---|
| Happy | 90.6 | Happy → Neutral | 6.5 |
| Sad | 88.0 | Sad → Neutral | 10.0 |
| Neutral | 96.3 | Neutral → Happy | 2.8 |
| Overall | 91.5 | — | — |

## B. Audio Preprocessing and Feature Extraction

Each audio sample was first resampled to a consistent rate of 16 kHz to standardize temporal resolution across all recordings, followed by amplitude normalization to a target level of −3 dBFS to ensure uniform loudness and prevent amplitude-related model bias. Preprocessing further included noise reduction, silence removal, and segmentation into frames using Hamming windows to ensure consistent spectral analysis conditions.

Following preprocessing, a 193-dimensional acoustic feature vector was extracted from each sample using the Librosa library. The feature set comprised five complementary representations chosen for their established relevance to emotion recognition: Mel-Frequency Cepstral Coefficients (MFCCs, 40 dimensions) capturing spectral envelope and vocal tract characteristics; Chroma features (12 dimensions) representing

the intensity of the 12 pitch classes and encoding harmonic content; Mel Spectrogram (128 dimensions) providing a time-frequency representation mapped to the perceptually motivated Mel scale according to $Mel(f) = 2595 \cdot \log_{10}(1 + f/700)$; Spectral Contrast (7 dimensions) measuring peak-to-valley differences across spectral sub-bands; and Tonnetz (6 dimensions) capturing tonal and harmonic relationships in the signal. Visual inspection of the extracted features confirmed distinct acoustic patterns across emotion classes, with Happy samples exhibiting higher energy and greater spectral variation, Sad samples showing lower intensity with compressed dynamic range, and Neutral samples displaying comparatively flat prosodic characteristics.

### C. Dimensionality Reduction and Encoding Preparation

The 193-dimensional feature vectors were first standardized using z-score normalization, where each feature i was transformed as $x'_i = (x_i - \mu_i) / \sigma_i$, with $\mu_i$ and $\sigma_i$ computed on the training set, ensuring equal contribution from all feature types regardless of their original scales. Principal Component Analysis (PCA) was then applied to the standardized feature matrix to reduce dimensionality, mitigate the curse of dimensionality, and remove redundant variance prior to quantum encoding.

Based on cumulative explained variance analysis, 27 principal components were retained, capturing approximately 80.6% of the total variance in the dataset. This component count was determined by the quantum circuit design: with 9 qubits and 3 features per qubit (as required by the SU(2) encoding described in Section 3.4), 27 components provided an exact mapping of all features to circuit parameters without padding or truncation. Prior to quantum circuit input, the PCA-reduced features were further scaled to the range $[0, \pi]$ using min-max normalization:

$$x'_i = \pi \cdot (x_i - x\_min) / (x\_max - x\_min) \qquad (1)$$

This scaling ensured compatibility with rotation gate parameters, which accept angles in radians.

### D. Variational Quantum Circuit

The classification model employed was a Variational Quantum Circuit, following the architecture introduced by Pérez-Salinas et al. [10] and the benchmarking implementation of Bowles et al. [14] The circuit operated on 9 qubits, with each qubit encoding 3 of the 27 PCA-reduced features per layer through a trainable SU(2) rotation block. For qubit $i$ and layer $l$, the encoding applied three consecutive rotation gates:

$$U\_enc(x_i, \omega^l_i, \theta^l_i) = R\_Z(x_2 \cdot \omega^3 + \theta^3) \cdot R\_Y(x_1 \cdot \omega^2 + \theta^2) \cdot R\_Z(x_0 \cdot \omega^1 + \theta^1) \qquad (2)$$

where $x_0$, $x_1$, $x_2$ are the three features assigned to qubit i, $\omega = (\omega^1, \omega^2, \omega^3)$ are trainable scaling weights, and $\theta = (\theta^1, \theta^2, \theta^3)$ are trainable bias parameters, both specific to layer l and qubit i. This formulation is the defining property of data re-uploading: unlike standard angle encoding, where features are embedded once into the circuit, the same

27 input features were *re-injected at every layer* with a distinct set of trainable (ω,θ) parameters, enabling the circuit to learn progressively richer Fourier representations of the input data across layers.

Following the encoding block at each layer, entanglement was introduced between adjacent qubits using a CZ ladder topology: CZ gates were applied sequentially between qubits (0,1), (1,2), ..., (7,8). This entanglement structure, combined with the repeated data injection, allows the circuit to capture complex inter-feature correlations that deepen with each additional layer. The complete circuit consisted of L such encoding-entanglement blocks, where L was treated as a hyperparameter.

Classification output was obtained through a multi-observable measurement (MORE) strategy on qubit 0, computing the expectation values of all three Pauli observables:

$$\langle O\_c \rangle = \langle \psi\_out | \, O\_c \, | \psi\_out \rangle, \; O\_c \in \{X_0, \, Y_0, \, Z_0\} \qquad (3)$$

yielding three scalar outputs corresponding to the three emotion classes (Happy, Sad, Neutral). The predicted class was determined by:

$$\hat{y} = argmax\_c \, \langle O\_c \rangle \qquad (4)$$

The model was trained by minimizing square loss between the measurement expectation values and the one-hot encoded class labels, following the loss formulation of Pérez-Salinas et al. [10]. The total number of trainable parameters for a circuit with L layers was L × 9 × 6, accounting for three ω and three θ values per qubit per layer.

## E. Training Configuration and Evaluation

The dataset was divided into 480 training samples (80%) and 120 test samples (20%), maintaining the balanced class distribution across both splits. Trainable parameters were initialized following the Pérez-Salinas convention: scaling weights $\omega \sim N(1.0, 0.1)$, reflecting near-unity pass-through scaling at initialization, and bias parameters $\theta \sim N(0.0, 0.1)$, reflecting near-zero bias. Training was terminated either upon reaching the 100-epoch budget or upon meeting the convergence criterion of Bowles et al.: training was stopped when the variance of the square loss over the most recent 200 consecutive parameter updates fell below a threshold, indicating that the optimizer had reached a stable minimum. All runs used a fixed random seed (42) for reproducibility, and performance was assessed on the held-out test set using per-class precision, recall, and F1-score, macro- and weighted-averaged F1-score, and a confusion matrix. Gradient norms were monitored throughout training as a diagnostic for the barren plateau phenomenon [15].

To identify a robust hyperparameter configuration without the compute cost of a full Cartesian sweep, a staged cross-ablation protocol was adopted, following the standard established in recent QML benchmarking work [14], [16], [17]. In Stage 1, a full grid was swept over the two most impactful axes — the number of variational layers $L \in \{1, 5, 10\}$ and the learning rate $\eta \in \{0.01, 0.1\}$ — yielding six core configurations,

with Adam as the optimizer and a batch size of 32. These values follow a log-scale coarse-search convention: the learning rates span one order of magnitude — the standard starting range for Adam-class optimizers on variational circuits and a direct subset of the $\eta \in \{0.001, 0.01, 0.1\}$ grid used in the QML benchmark of Bowles et al. [14] — while the depth values cover the minimum non-trivial re-uploading layer (L = 1), moderate depth (L = 5), and the upper bound at which barren-plateau risk is conventionally expected to emerge (L = 10) [15], matching the depth range investigated by Pérez-Salinas et al. [10] in the original data-reuploading paper. In Stage 2, the core optimum (L*, η*) was fixed, and three optimizers were compared: Adam (the Stage 1 default), RMSprop, and stochastic gradient descent with Nesterov momentum, matching the selection of Moussa et al. [16] and Herbst et al. [17], who identify optimizer choice as among the two most important hyperparameters for quantum neural network classification. Stage 3 fixed (L*, η*, optimizer*) and compared batch sizes {16, 32, 64} following the log-2 sweep convention of the Deep Learning Tuning Playbook [18]. The best configuration identified at each stage was carried forward to the next; the final configuration served as the expanded best model for the detailed analysis in Section 4.1.3 and Section 4.1.4.

### F. Classical Baselines and Speaker-Independent Evaluation

To contextualize the performance of the proposed VQC, two classical baselines were trained and evaluated on the same feature pipeline and data split. The Random Forest (RF) is a non-parametric ensemble that has been shown to outperform deep neural architectures on low-resource SER corpora where limited training data makes convolutional and recurrent models prone to overfitting [4]; it therefore serves as the strong non-neural classical benchmark. The Multilayer Perceptron (MLP) with a single hidden layer is the closest classical analogue to the VQC in structural terms, as both are parametric function approximators with tunable weights trained by gradient descent on a differentiable loss; this allows a direct comparison of representational capacity between the two parametric paradigms. Convolutional and recurrent architectures were deliberately excluded from the comparison, as the 27-dimensional PCA-reduced feature vector lacks the sequential or spatial structure these models are designed to exploit.

Both classical models were trained on the identical PCA(27) feature space and the identical 480/120 stratified split used for the VQC, with random_state=42 throughout. Per-model hyperparameters were tuned by five-fold cross-validation on the 480 training samples only; the held-out 120-sample test set was used for a single final evaluation per model, matching the single-run protocol used for the VQC. Hyperparameter grids were: Random Forest over n_estimators $\in\{100, 300, 500\}\times$max_depth $\in\{$None,10,20$\}$; and MLP over hidden_layer_sizes $\in\{32,64,128\}\times$learning_rate_init $\in\{0.001, 0.01\}$, trained for up to 500 epochs with early stopping on validation loss.

The stratified 480/120 split used throughout this study places samples from all ten speakers in both the training and the test set; no speaker is held out entirely. Under this protocol, a classifier that learns speaker-identifying acoustic features — timbre, fundamental-frequency range, formant positions — can exploit them as shortcuts for emotion classification, because the same voices appear on both sides of the split. To quantify the extent of this leakage on the classical baselines, both Random Forest and MLP were additionally evaluated under a leave-one-speaker-out (LOSO) cross-validation protocol. In each of the ten folds, the classifier is trained on nine speakers (540 samples) and tested on the remaining speaker (60 samples); the classifier never sees the test speaker's voice during training. Hyperparameters were re-tuned per fold via the same five-fold inner cross-validation as in the speaker-dependent protocol above. Results of both evaluations are reported in Section 4.2.

## 4. RESULTS

### A. Performance of the Variational Quantum Circuit

#### a. Hyperparameter Grid Search

The hyperparameter grid search was conducted on the VQC classifier across the core axes of circuit depth and learning rate. Six configurations were trained and evaluated on the held-out test set: three circuit depths $L \in \{1, 5, 10\}$ × two learning rates $\eta \in \{0.01, 0.1\}$. All runs used the Adam optimizer with a batch size of 32 and Pérez-Salinas initialization, with convergence detected via the loss-variance criterion described in Section 3.5. Results are summarized in Table 4, and Figure 3 visualizes the accuracy and macro-F1 surfaces across the grid.

The grid clearly identifies a dominant configuration at $L = 10$, $\eta = 0.01$, which attains 88.33% test accuracy and a macro-F1 of 0.883. Two structural observations follow from the grid. First, the lower learning rate ($\eta = 0.01$) outperforms the higher one ($\eta = 0.1$) at every depth, consistent with the expectation that smaller step sizes are required to avoid overshoot in rugged variational loss landscapes. Second, test accuracy increases monotonically with depth but at a diminishing rate — the gain from L=1 to L=5 is substantial, while the gain from L=5 to L=10 is more modest. This saturation pattern suggests that the representational capacity of the data-reuploading architecture has largely been recruited by depth 10 on the present feature pipeline. Figure 3 visualizes the accuracy and F1-score surfaces across the grid, confirming the dominance of lower learning rates and the saturation of performance gains beyond moderate circuit depth.

**Table 4**: Hyperparameter grid search results. Best configuration highlighted.

| L | η | Params | Train Acc (%) | Test Acc (%) | F1-Score | Converged |
|---|---|---|---|---|---|---|
| 1 | 0.01 | 54 | 55.0 | 62.5 | 0.622 | No |
| 1 | 0.10 | 54 | 56.7 | 63.3 | 0.616 | Yes |
| 5 | 0.01 | 270 | 83.8 | 86.7 | 0.868 | Yes |
| 5 | 0.10 | 270 | 78.5 | 81.7 | 0.817 | Yes |
| 10 | 0.01 | 540 | 90.0 | 88.3 | 0.883 | No |
| 10 | 0.10 | 540 | 82.1 | 81.7 | 0.812 | Yes |

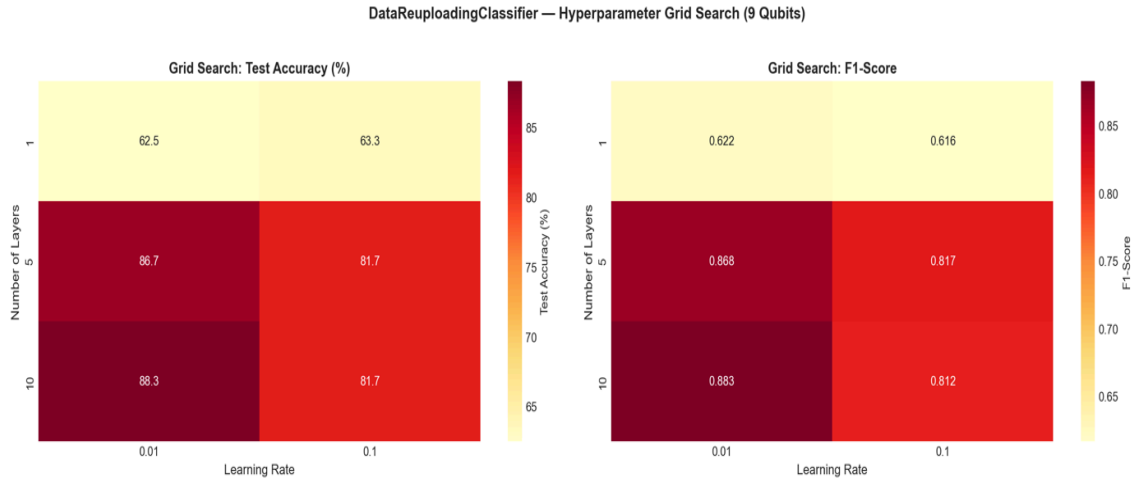


**Figure 3**: Hyperparameter grid search over circuit depth L ∈ {1, 5, 10} and learning rate η ∈ {0.01, 0.1}. Left: test accuracy (%). Right: macro F1-score. Best configuration: L = 10, η = 0.01.

### b. Extended Hyperparameter Exploration

Building on the observations of Section 4.1.1, the remaining hyperparameter axes were explored following the staged ablation protocol described in Section 3.5. Starting from the core optimum (L = 10, η = 0.01, Adam, batch = 32), the optimizer and batch size were ablated individually, with each axis anchored on the best configuration identified in the preceding stage. Results for the six additional configurations are summarized in Table 5.

**Table 5**: Extended hyperparameter exploration. Two ablation axes — optimizer and batch size — anchored on the best core configuration from Section 4.1.1 (L=10, η=0.01). Anchors indicated by "(anchor)"; the configuration carried into the next stage highlighted in bold

| Ablation axis | Configuration | Test accuracy | Macro-F1 |
|---|---|---|---|
| Optimizer | Adam (anchor) | 88.33% | 0.883 |
| Optimizer | RMSprop | 87.50% | 0.877 |
| Optimizer | SGD + Nesterov momentum | 81.67% | 0.817 |
| Batch size | 16 | 89.17% | 0.891 |
| Batch size | 32 (anchor) | 88.33% | 0.883 |
| **Batch size** | 64 ★ | **90.83%** | **0.908** |

The optimizer ablation produced a modest spread of 6.66 percentage points. Adam remained the best choice at 88.33%, with RMSprop close behind at 87.50% (F1=0.877); stochastic gradient descent with Nesterov momentum lagged notably at 81.67%, consistent with the broader observation in the QML benchmarking literature that gradient-based adaptive methods outperform momentum-only SGD on variational quantum models. The batch-size ablation yielded the most consequential improvement in the extended search: batch size 64 lifted test accuracy to 90.83% (F1=0.908), a 2.50 percentage-point gain over the batch-32 anchor.

The best configuration across all stages — L = 10, η = 0.01, Adam, batch = 64 — attains 90.83% test accuracy and macro-F1=0.9082, an improvement of 2.50 percentage points over the core-grid optimum in Section 4.1.1. This configuration is used as the best model for the detailed analysis in Sections 4.1.3 and 4.1.4, and as the VQC comparator in the classical-baselines comparison in Section 4.2.

### c. Training Dynamics of the Best Model

The training dynamics of the best configuration (L=10, η=0.01, Adam, batch=64, Pérez-Salinas initialization) over 100 epochs are presented in Figure 5. Test accuracy exhibited a characteristic two-phase profile: a rapid increase during the first 20–30 epochs as the circuit learned coarse class separability, followed by a slower fine-tuning phase thereafter (as shown in Figure 4). The best test accuracy of 90.83% was recorded at epoch 48 via checkpoint selection.

The gradient norm $\|\nabla L\|$ remained consistently active across all 100 epochs, with a mean of 0.7403 and a range of [0.3342, 1.5894]. The ratio between the smallest and

largest values ($\log_{10}$ ratio ≈ 0.68) indicates that no epoch experienced pathological gradient concentration. This directly confirms the absence of barren plateaus at the reported depth. Ten-layer parameterized quantum circuits are conventionally expected to exhibit vanishing-gradient phenomena due to the exponential concentration of the loss landscape with increasing system size; the data-reuploading architecture with trainable SU(2) encoding appears to mitigate this effect at nine qubits and L = 10, preserving an exploitable gradient signal throughout training.

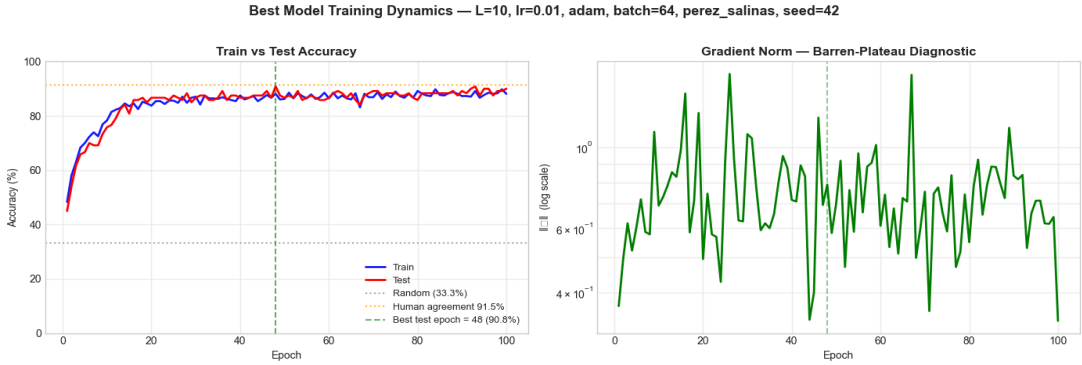


**Figure 4**: Training dynamics of the best model (L =10, η=0.01, Adam, batch=64, Pérez-Salinas, 540 parameters). (a) Train and test accuracy over 100 epochs, with the best epoch (48) marked in red. (b) Gradient norm $\|\nabla L\|$ on a logarithmic scale, confirming the absence of barren plateaus throughout training (mean 0.7403, range [0.3342, 1.5894])

### d. Classification Performance

Table 6 presents per-class precision, recall, and F1-score for the best VQC model on the held-out test set of 120 samples (40 per class). Figure 5 shows the corresponding confusion matrices in raw counts and row-normalized recall form.

**Table 6**: Per-class classification metrics for the best VQC model (L=10, η=0.01, Adam, batch= 64, Pérez-Salinas)

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Happy | 0.925 | 0.925 | 0.925 | 40 |
| Sad | 0.902 | 0.925 | 0.914 | 40 |
| Neutral | 0.897 | 0.875 | 0.886 | 40 |
| **Macro average** | **0.908** | **0.908** | **0.908** | **120** |
| Weighted average | 0.908 | 0.908 | 0.908 | 120 |

The classifier achieves balanced performance across all three emotion classes, with macro-averaged precision, recall, and F1-score all equal to 0.908. Happy is recognized most consistently (precision and recall both 0.925), with only three misclassifications out of forty: two samples assigned to Sad and one to Neutral. Sad attains the same recall (0.925) with three misclassifications to Neutral, reflecting a modest acoustic overlap between low-arousal states. Neutral is classified slightly below the other two classes (precision 0.897, recall 0.875), with five misclassifications distributed between Happy (three samples) and Sad (two samples); this Neutral-to-Happy confusion likely reflects the acoustic proximity between mildly expressed happiness and neutral prosody in Nepali speech, where tonal variation is subtler than in tonal or high-resource languages. Overall, no confusion rate exceeds 7.5% in any direction, and the confusion matrix is close to diagonally dominant, indicating that the best VQC model captures emotion-discriminative structure in the PCA-reduced feature space without systematic bias toward any class.

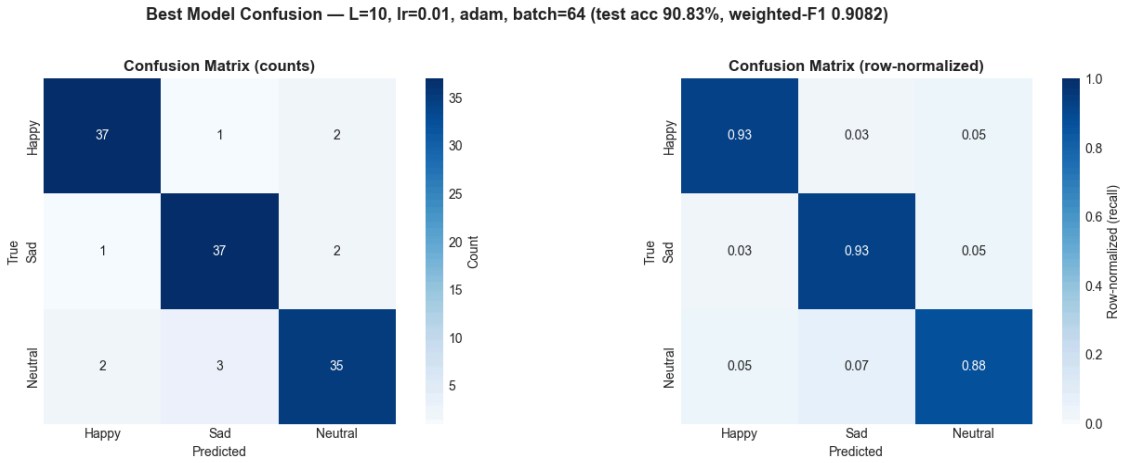


**Figure 5**: Confusion matrices for the best VQC model (L=10, η=0.01, Adam, batch=64). (a) Raw counts, n=120. (b) Row-normalized recall rates. The largest off-diagonal mass corresponds to Neutral → Happy (7.5%).

## B. Classical Baselines Comparison

Following the protocol described in Section 3.6, two classical baselines — Random Forest and a single-hidden-layer MLP — were evaluated on the identical feature pipeline and data split used for the VQC. Table 7 reports the test accuracy, macro-F1, and best hyperparameters for both baselines alongside the proposed VQC.

**Table 7**: Classical baselines vs. the proposed VQC. All models on PCA(27) features and the identical 480/120 stratified split (random_state= 42). Inner 5-fold CV on the training set only; single evaluation on the held-out test set.

| Model | Best hyperparameters | Params | Test Accuracy | Macro-F1 |
|---|---|---|---|---|
| Random Forest | n_estimators = 100, max_depth = None | — | 95.00% | 0.950 |
| MLP (1 hidden layer) | hidden = (128,), lr = 0.001 | ~3,840 | 99.17% | 0.992 |
| **VQC (proposed, best)** | **L = 10, η = 0.01, Adam, batch = 64, Pérez-Salinas** | **540** | **90.83%** | **0.908** |

**Table 8**: Per-class classification metrics for the classical baselines (Random Forest and MLP) on the held-out test set of 120 samples (40 per class). Evaluated at random_state=42, matching the VQC protocol.

| Model | Class | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Random Forest | Happy | 0.975 | 0.975 | 0.975 |
| | Sad | 0.974 | 0.925 | 0.949 |
| | Neutral | 0.905 | 0.950 | 0.927 |
| | **Macro avg** | **0.951** | **0.950** | **0.950** |
| MLP | Happy | 1.000 | 0.975 | 0.987 |
| | Sad | 1.000 | 1.000 | 1.000 |
| | Neutral | 0.976 | 1.000 | 0.988 |
| | **Macro avg** | **0.992** | **0.992** | **0.992** |

Table 8 provides the per-class precision, recall, and F1-score for both classical baselines on the same held-out test set. Both classical baselines outperform the VQC on this held-out test set, with Random Forest reaching 95.00% accuracy and MLP reaching 99.17%, compared to 90.83% for the VQC at its expanded best configuration. The margins are 4.17 and 8.34 percentage points respectively. The MLP attained near-perfect performance across all three emotion classes, while the Random Forest showed a slightly weaker Neutral-class precision (0.905) balanced by a higher Neutral recall (0.950), consistent with its inductive bias toward the class that occupies the most

heterogeneous region of the feature space. The proposed VQC is therefore not competitive with tuned classical baselines under this speaker-dependent protocol; however, as Table 9 shows, a substantial portion of the classical advantage on this corpus reflects speaker-level information leakage rather than emotion-discriminative learning.

**Table 9**: Speaker-independent evaluation of classical baselines via leave-one-speaker-out (LOSO) cross-validation. Mean and standard deviation of fold-level test accuracy across ten held-out speakers; pooled out-of-fold accuracy computed across all 600 predictions.

| Model | LOSO Acc (mean ± std) | 95% CI | Pooled OOF | Δ vs speaker-dep |
|---|---|---|---|---|
| Random Forest | 61.50 ± 13.34% | [53.50, 69.00] | 61.50% | −33.50 pp |
| MLP | 66.67 ± 9.23% | [60.16, 71.50] | 66.67% | −32.50 pp |

Table 9 reports the speaker-independent evaluation of both classical baselines under the leave-one-speaker-out (LOSO) protocol described in Section 3.6. Both classical baselines collapse substantially under speaker-independent evaluation. Random Forest accuracy drops from 95.00% to 61.50% (a 33.50 percentage-point decline), and MLP accuracy drops from 99.17% to 66.67% (a 32.50 percentage-point decline). Put differently, approximately one-third of each classical model's headline test accuracy on this corpus is attributable to speaker leakage rather than to emotion-discriminative learning. The wide per-speaker standard deviations (13.34 pp for RF, 9.23 pp for MLP) further indicate that classical performance is highly speaker-dependent even under leak-free evaluation.

A matched LOSO evaluation of the VQC is computationally prohibitive at its current 540-parameter, 100-epoch training configuration and is deferred to future work. However, a parameter-count comparison is informative. The MLP that attains 99.17% under the leaky protocol and 66.67% under LOSO is parameterized with approximately 3,840 trainable weights, roughly seven times the VQC's 540. Recent work on the generalization properties of variational quantum models has documented that VQCs can exhibit reduced overfitting compared to classical neural networks of comparable capacity on small datasets [16], [17], a property especially relevant in the small-sample regime of the present corpus. Whether this general property extends specifically to reduced susceptibility to speaker-shortcut learning cannot be determined from the present experiments and remains an open empirical question; nonetheless, the VQC's substantially smaller parameter footprint suggests that its headline 90.83% test accuracy may be less inflated by speaker-identity shortcuts than the MLP's 99.17% at roughly seven times the parameter count.

## 5. DISCUSSION

### A. Principal Findings

The best VQC configuration identified through the staged cross-ablation — L=10, η=0.01, Adam, batch size=64, Pérez-Salinas initialization — attains 90.83% test accuracy and macro-F1= 0.908 on the Nepali-SER corpus. Three findings from the Results section bear further interpretation.

The staged hyperparameter exploration identified a configuration (batch size 64) that improved test accuracy by 2.50 percentage points over the core-grid optimum (Table 5). This improvement, obtained without any change to the core architecture, suggests that the batch-size axis is underexplored in the current QML benchmarking literature — most recent VQC studies on small corpora default to batch sizes of 16 or 32, consistent with simulation-cost constraints, but do not systematically explore larger batches. The gain observed here is consistent with the broader machine-learning finding that larger batch sizes can provide lower-variance gradient estimates on small training sets, though the generalization behaviour of this phenomenon on variational circuits deserves a dedicated future study.

The gradient-norm analysis confirms the absence of barren plateaus across all 100 training epochs at depth L=10, with mean gradient norm 0.7403 and range [0.3342, 1.5894]. This is a non-trivial methodological result: conventional analyses of parameterized quantum circuits predict that expressivity at this depth is accompanied by exponential gradient concentration [15], and the fact that the data-reuploading architecture with trainable SU(2) encoding avoids this pathology on the present feature pipeline merits independent investigation. The trainable encoding — in which both the data-scaling weights ω and the bias parameters θ are learned rather than fixed — appears to break the symmetry assumptions under which barren plateaus arise, though a formal theoretical characterization lies outside the scope of this paper.

Comparison with classical baselines (Table 7) shows that tuned Random Forest and MLP models achieve higher test accuracy than the VQC on the speaker-dependent protocol, by margins of 4.17 and 8.34 percentage points respectively. The proposed VQC is therefore not competitive with well-tuned classical methods on this protocol. However, the speaker-independent robustness check (Table 9) reveals that both classical baselines collapse to 61.50% (RF) and 66.67% (MLP) under leave-one-speaker-out evaluation — accuracy drops of approximately one-third. This indicates that a substantial portion of the classical advantage on this corpus reflects the model's exploitation of speaker-identifying acoustic features rather than genuinely emotion-discriminative learning. The parameter-count gap between the models is also pertinent: the MLP achieves 99.17% with approximately 3,840 trainable weights, roughly seven times the VQC's 540, and prior work on the generalization properties of variational quantum models has documented that VQCs can exhibit reduced overfitting compared to classical neural networks of comparable capacity on small datasets [16],

[17]. Whether this general property extends specifically to reduced susceptibility to speaker-shortcut learning cannot be determined from the present experiments and remains an open empirical question; the VQC's substantially smaller parameter footprint suggests, but does not establish, that its 90.83% accuracy may be less inflated by speaker-identity shortcuts than the MLP's 99.17%.

### B. Limitations

Four limitations of the present study constrain the interpretation of its results. First, the Nepali-SER corpus comprises 600 recordings from 10 male native speakers across three emotion classes, and this scale imposes a strict ceiling on the generalization claims that can be made from any classifier trained on it. The balanced class distribution and perceptual validation at 91.5% mitigate but do not eliminate the small-sample concern. Second, the 9-qubit, 27-component PCA pipeline was dictated by the SU(2) encoding's three-features-per-qubit capacity, which may handicap the quantum model relative to pipelines with different feature-selection criteria — in particular, methods such as mutual-information feature selection are not constrained by qubit budgets and may capture task-relevant dimensions that PCA does not. Third, all results reported in this study are from noise-free analytic simulation; on physical NISQ hardware, gate errors, decoherence, and shot-based readout noise are expected to degrade classifier accuracy further, with the magnitude of degradation dependent on the specific hardware platform and error-mitigation strategy [13]. Fourth, a matched leave-one-speaker-out evaluation of the proposed VQC was not feasible within the compute budget of the present study; the speaker-independence comparison in section 4.2 is therefore indicative rather than conclusive, and a fully matched LOSO comparison is left to the extension noted in section 6.

## 6. SUGGESTIONS AND RECOMMENDATIONS

Future work should pursue four directions. First, the Nepali emotional-speech corpus should be expanded beyond three emotion classes and 600 utterances, with additional classes drawn from the canonical set — anger, fear, surprise, disgust — and broader speaker diversity along gender, age, and regional-dialect axes. While 200 recordings per emotion across 10 actors is adequate to characterise each emotion for this pilot study, an expanded corpus with more speakers and additional emotional categories is needed to establish generalisable Nepali SER performance. Second, a matched leave-one-speaker-out evaluation of the proposed VQC is a direct follow-up that would resolve the methodological asymmetry in the present comparison; the ten fold-specific trainings required are modest in absolute terms and are planned for an extension of this study. Third, hybrid quantum–classical architectures — in which a parameterized quantum circuit serves as a feature-extraction layer within a larger classical network, as in Rajapakshe et al. [12] — should be compared against both the pure VQC evaluated here

and the classical baselines; prior work suggests that hybrid designs can be competitive where pure VQCs are not, and the parameter-efficiency argument from section 5.1 may be further strengthened in hybrid regimes. Fourth, noise-resilient encoding strategies and error-mitigation techniques should be investigated before any deployment on physical NISQ hardware; alternative feature-selection pipelines — particularly mutual-information-based methods — may also close part of the accuracy gap between quantum and classical models on this corpus and are worth exploring in parallel.

## ACKNOWLEDGEMENTS

The authors are genuinely grateful to the voice actors who contributed to this dataset. We also thank the emotion validators who took part in the perceptual validation.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are not publicly available but are available from the author on reasonable request.

## REFERENCES

[1] K. R. Scherer, "Vocal affect expression: A review and a model for future research,"Psychological Bulletin, vol. 99, no. 2, pp. 143–165, 1986.

[2] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," Sensors, vol. 21, no. 4, Art. no. 1249, 2021.

[3] A. Monisha, "A review of the advancement in speech emotion recognition for Indo-Aryan and Dravidian languages," Advances in Human-Computer Interaction, vol. 2022, Art. no. 9602429, 2022.

[4] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," PLoS ONE, vol. 18, no. 11, Art. no. e0291500, 2023.

[5] M. Cerezo et al., "Variational quantum algorithms," Nature Reviews Physics, vol. 3, no. 9, pp. 625–644, 2021.

[6] M. C. Caro et al., "Generalization in quantum machine learning from few training data," Nature Communications, vol. 13, Art. no. 4919, 2022.

[7] Y. Wang and A. Boumadane, "A fine-tuned wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," arXiv preprint arXiv:2111.02735, 2022.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Proc. Interspeech, Lisbon, Portugal, 2005, pp. 1517–1520.

[9] M. Schuld, R. Sweke, and J. J. Meyer, "Effect of data encoding on the expressive power of variational quantum-machine-learning models," Physical Review A, vol. 103, no. 3, Art. no. 032430, 2021.

[10] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, "Data re-uploading for a universal quantum classifier," Quantum, vol. 4, p. 226, 2020.

[11] M. Rath and H. Date, "Quantum data encoding: A comparative analysis of classical-to-quantum mapping techniques and their impact on machine learning accuracy," EPJ Quantum Technology, vol. 11, Art. no. 72, 2024.

[12] T. Rajapakshe, R. Rana, F. Riaz, S. Khalifa, and B. W. Schuller, "Representation learning with parameterised quantum circuits for advancing speech emotion recognition," Scientific Reports, vol. 15, Art. no. 44353, 2025.

[13] R. Norval and D. Wang, "Quantum AI in speech emotion recognition," Entropy, vol. 27, no. 12, Art. no. 1201, 2025.

[14] J. Bowles, S. Ahmed, and M. Schuld, "Better than classical? The subtle art of benchmarking quantum machine learning models," arXiv preprint arXiv:2403.07059, 2024.

[15] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nature Communications*, vol. 9, Art. no. 4812, 2018.

[16] J. J. Moussa, J. N. van Rijn, T. Bäck, and V. Dunjko, "Hyperparameter importance of quantum neural networks across small datasets," Machine Learning, 2023; arXiv:2206.09992.

[17] M. Herbst, V. De Maio, and I. Brandic, "On optimizing hyperparameters for quantum neural networks," in Proc. IEEE Int. Conf. Quantum Computing and Engineering (QCE), 2024; arXiv:2403.18579.

[18] V. Godbole, G. E. Dahl, J. Gilmer, C. J. Shallue, and Z. Nado, "Deep Learning Tuning Playbook," GitHub repository, 2023. Available: github.com/google-research/tuning_playbook.