SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Hybrid Quantum Re-Uploading Framework for Multimodal Emotion Recognition

Sai Siva Naga Chandra Pathuru[1*], S. Sri Harsha[2]

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (K L E F), Vaddeswaram, Guntur, Andhra Pradesh, India. E-mail: 2401050035@kluniversity.in (S.S.N.C.P.)
[2] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (K L E F), Vaddeswaram, Guntur, Andhra Pradesh, India. E-mail: sharsha@kluniversity.in (S.S.H.)

Abstract
Multimodal emotion recognition can benefit from combining complementary information from facial images, speech, and text. However, integrating heterogeneous modalities obtained from independently collected datasets introduces challenges in data alignment and label consistency. This study presents a hybrid quantum–classical framework that combines modality-specific classical encoders with a quantum re-uploading variational classifier for six-class emotion recognition. Facial images, speech samples, and textual samples were obtained from separate publicly available datasets and aligned at the emotion-label level to construct a unified multimodal experimental setting. Image, speech, and text encoders generated modality-specific representations that were fused through a classical neural network and projected into a six-dimensional quantum input space. A six-qubit quantum circuit with three re-uploading layers repeatedly encoded the fused features before six-class classification. On a test set of 6,347 samples, the classical Image–Speech–Text fusion model achieved 73.94% accuracy and a 70.64% macro F1-score, whereas the proposed quantum re-uploading model achieved 72.57% accuracy and a 68.76% macro F1-score. The quantum model used 8,706 trainable parameters compared with 15,440,006 for the classical multimodal model, representing a 99.94% reduction in trainable parameters. The results indicate that quantum re-uploading can provide competitive performance for multimodal emotion recognition while highlighting the limitations of label-level alignment across independently collected datasets.

Keywords: Quantum Machine Learning; Multimodal Emotion Recognition; Quantum Re-Uploading; Affective Computing; Hybrid Quantum-Classical Models.

1. Introduction

Emotions of human beings are at the center of thinking, making choices, communication and interaction with others. Emotional state recognition and interpretation have thus found application as a core requirement to many intelligent systems, such as human-computer interaction systems, affect-aware recommender systems, virtual assistants, and mental health monitoring systems [1]. Emotion recognition is used to approximate the difference between the affective behavior of a human being and that of a machine by modeling the expression and perception of emotions using observable cues.

1.1 Evolution from Unimodal to Multimodal Approaches

The initial studies of emotion recognition have mainly been based on unimodal data sources like facial expression, speech, or physiological datum [2]. Although the unimodal methods showed good performance in the controlled environment, their effectiveness deteriorated in the actual environment because of noise, ambiguity and inter-subject differences [3]. Human emotions are complex, situation-dependent, and ambiguous and one modality is usually not sufficient to represent the entire range of affective information [4].

As a solution to these shortcomings, multimodal emotion recognition has become a mainstream line of research [5, 6]. Multimodal systems exploit complementary information across heterogeneous data sources, including speech, text, and facial images [7–9]. In this study, these three modalities are considered because they provide complementary acoustic, semantic, and visual information while allowing a unified six-class experimental setting through label-level alignment of samples obtained from separate datasets. Speech provides prosodic and acoustic information, text provides semantic and contextual information, and facial expressions provide visible affective cues [10, 11]. The combination of these modalities can provide more comprehensive affective information than individual modalities alone [4, 12, 13]. This complementary nature motivates the use of multimodal feature fusion as the classical foundation of the framework developed in this study.

1.2 Deep Learning Paradigms in Emotion Recognition

The recent developments in deep learning have greatly enhanced the evolution in relation to multimodal emotion recognition. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based

models have been demonstrated very popular in learning modality-specific representations and feature-level/ decision-level fusion [7, 9, 14]. The fusion strategies that have been studied extensively to integrate information across modalities are multimodal fusion strategies which include early fusion, late fusion and hybrid fusion [5, 13]. More recent systems utilize the mechanisms of attention, and large multimodal language models as dynamic ways of modeling cross-modal interactions [11, 14–16]. Recent attention-based and graph-based architectures demonstrate that effective multimodal emotion recognition depends on learning interactions between modality-specific representations rather than simply concatenating independent features [11, 16]. These developments motivate the use of modality-specific encoders followed by feature fusion in the present framework, while avoiding the computational complexity of fully transformer-based multimodal architectures. Although effective, classical deep learning models can require substantial computational resources and large labelled datasets, particularly as multimodal architectures become increasingly complex [6, 13, 17]. These limitations motivate investigating alternative computational paradigms that can retain the benefits of learned multimodal representations while introducing more compact trainable components. Quantum machine learning provides one such direction.

1.3 Quantum Machine Learning: A Paradigm Shift

Simultaneously with classical machine learning, quantum computing has become a potential paradigm to solve complex learning problems [14, 15]. Quantum machine learning (QML) studies the combination of quantum computing concepts and machine learning algorithms to use quantum phenomena including superposition and entanglement to improve the capacity to represent [16, 17]. The use of hybrid quantum-classical models has been of especial interest to near-term noisy intermediate-scale quantum (NISQ) devices [18, 19].

The core of several QML methods consists of variational quantum circuits (VQCs), and they can be optimized using classical methods [14, 16]. Several studies have explored quantum learning for emotion-related tasks, including speech emotion recognition, EEG-based emotion detection, facial emotion recognition, and multimodal learning [1–3, 19–21]. Collectively, these studies indicate that quantum models can be integrated with classical feature extraction and fusion pipelines, making hybrid quantum–classical learning particularly relevant for heterogeneous emotion data. However, much of the existing work focuses on individual modalities or conventional variational quantum circuits, leaving scope for investigating quantum re-uploading after multimodal classical representation learning.

Data re-uploading provides a mechanism for repeatedly encoding classical features across successive quantum layers, allowing the interaction between the input data and trainable quantum operations to be increased within a compact circuit [22]. The data re-uploading approach also establishes a trade-off between the number of qubits and the number of data-encoding layers, providing a resource-efficient strategy for quantum classification [22]. Recent studies have demonstrated that quantum re-uploading can improve learning performance while using relatively small quantum resources [22]. These properties motivate the architecture proposed in this study: classical encoders first learn modality-specific representations, classical fusion integrates complementary information from the three modalities, and a compact quantum re-uploading circuit subsequently performs six-class classification.

1.4 Related Work

1.4.1 Quantum-Inspired Classical Approaches

Quantum-inspired classical approaches apply mathematical concepts associated with quantum theory to classical machine-learning models without requiring quantum hardware [23]. In multimodal emotion and sentiment analysis, such approaches have been explored to represent uncertainty, modality interactions, and complex decision relationships [23–27]. For example, quantum-inspired adaptive-priority learning and related models demonstrate how quantum-inspired representations can be incorporated into multimodal emotion recognition using conventional computational platforms [24, 27]. These studies show that quantum-inspired principles can improve the representation of ambiguous and heterogeneous affective information while remaining compatible with classical learning systems.

1.4.2 Cognitive and Appraisal-Based Quantum Models

Cognitive and appraisal-based theories emphasize that emotional responses depend not only on observable stimuli but also on contextual interpretation, uncertainty, and subjective evaluation [10, 17]. Quantum cognition provides a mathematical framework for representing some of these uncertain and context-dependent decision processes [26]. Recent quantum-cognitive and quantum-inspired models have therefore investigated uncertainty-aware decision and fusion mechanisms for sentiment and emotion analysis [23–27]. Although these approaches are conceptually related to affective interpretation, they primarily motivate the representation of uncertainty and ambiguity rather than directly defining the architecture used in the present study.

1.4.3 Recent Advancements in Multimodal Emotion Recognition

Recent multimodal emotion recognition research has increasingly focused on improving robustness, scalability, and generalization under realistic conditions. Semi-supervised and open-vocabulary approaches address limitations associated with scarce labeled data and fixed emotion categories [9, 12], while recent Transformer and mixture-of-experts architectures improve the modeling of cross-modal and conversational interactions [8, 13, 28,

29]. New datasets have also expanded evaluation to settings such as virtual reality, educational environments, and standardized multimodal emotion perception [10, 11, 30]. Federated learning has further investigated privacy-preserving training across distributed data sources [15]. Collectively, these developments demonstrate that current multimodal emotion recognition research is moving toward more flexible architectures, richer datasets, and realistic deployment settings [14, 18, 19]. However, these advances remain predominantly within classical deep-learning paradigms, leaving an opportunity to investigate whether compact quantum components can complement established multimodal representation-learning pipelines.

1.5 Research Gap and Contributions

Although multimodal emotion recognition and quantum machine learning have each advanced considerably, their integration remains relatively limited. Existing multimodal emotion recognition studies have primarily focused on classical deep-learning architectures for learning cross-modal representations [14, 17, 18, 28, 29], while quantum emotion-recognition studies have mainly investigated individual modalities or specific quantum learning configurations [1–3, 20, 21, 23, 25]. Quantum-inspired and quantum-cognitive approaches have further explored uncertainty, ambiguity, and decision processes in affective analysis [23–27], but these approaches do not directly examine the use of a compact quantum re-uploading classifier after classical multimodal feature fusion. This study addresses this gap by combining established modality-specific classical encoders with a quantum re-uploading classifier within a unified three-modality framework.

The main contributions of this study are as follows:

1. A hybrid quantum–classical framework is developed for six-class emotion recognition using facial images, speech, and text. The three modalities are obtained from separate datasets and aligned at the emotion-label level, providing a unified experimental setting while explicitly acknowledging the limitations of cross-dataset alignment.
2. A classical multimodal fusion pipeline is established in which modality-specific encoders generate image, speech, and text representations that are integrated before quantum processing. Unimodal and pairwise configurations are also evaluated to examine the contribution of multimodal fusion.
3. A six-qubit quantum re-uploading classifier with three data re-uploading layers is integrated after classical feature fusion. The design uses repeated feature encoding to provide a compact quantum classification component [22].
4. The proposed quantum model is experimentally compared with the corresponding classical Image–Speech–Text fusion model using test accuracy, macro F1-score, per-class performance, and confusion-matrix analysis. The evaluation demonstrates that the quantum component achieves competitive performance close to the classical multimodal baseline, although it does not exceed the classical model.
5. The study analyzes the parameter efficiency of the two approaches and discusses the practical limitations associated with cross-dataset label-level alignment and the use of simulated quantum circuits.

2. Materials and Methods

2.1 Overall Architecture

The proposed framework is a hybrid quantum–classical architecture for six-class multimodal emotion recognition using facial images, speech, and text. The three modalities are obtained from separate publicly available datasets and are aligned at the emotion-label level before multimodal fusion. Each modality is processed by a dedicated classical encoder to obtain a modality-specific representation. The resulting representations are concatenated and passed through a classical fusion network to obtain a compact multimodal representation. This fused representation is subsequently projected into a six-dimensional quantum input space and processed by a six-qubit quantum re-uploading circuit for final emotion classification.

The classical encoders and fusion network are trained as part of the classical multimodal model. The best classical model is selected according to validation accuracy, while macro F1-score is monitored as an additional validation metric. For the quantum experiment, the trained three-modality encoders and fusion network are kept fixed, and their 128-dimensional fused representations are reused as input to the quantum re-uploading classifier. The quantum classifier is trained separately on the resulting fused representations, and its best checkpoint is selected according to validation accuracy. The validation set is therefore used for model selection and hyperparameter monitoring, while the test set is reserved for final performance evaluation.

The final architecture consists of three modality-specific encoders, a classical fusion stage, a projection layer that maps the 128-dimensional fused representation to six quantum inputs, and a six-qubit quantum re-uploading classifier with three re-uploading layers. The quantum circuit repeatedly encodes the projected features, applies trainable quantum rotations and nearest-neighbor entanglement, and measures the Pauli-Z expectation values. These six expectation values are passed to a classical classification layer to obtain the six emotion logits. The complete information flow is illustrated in Figure 1.

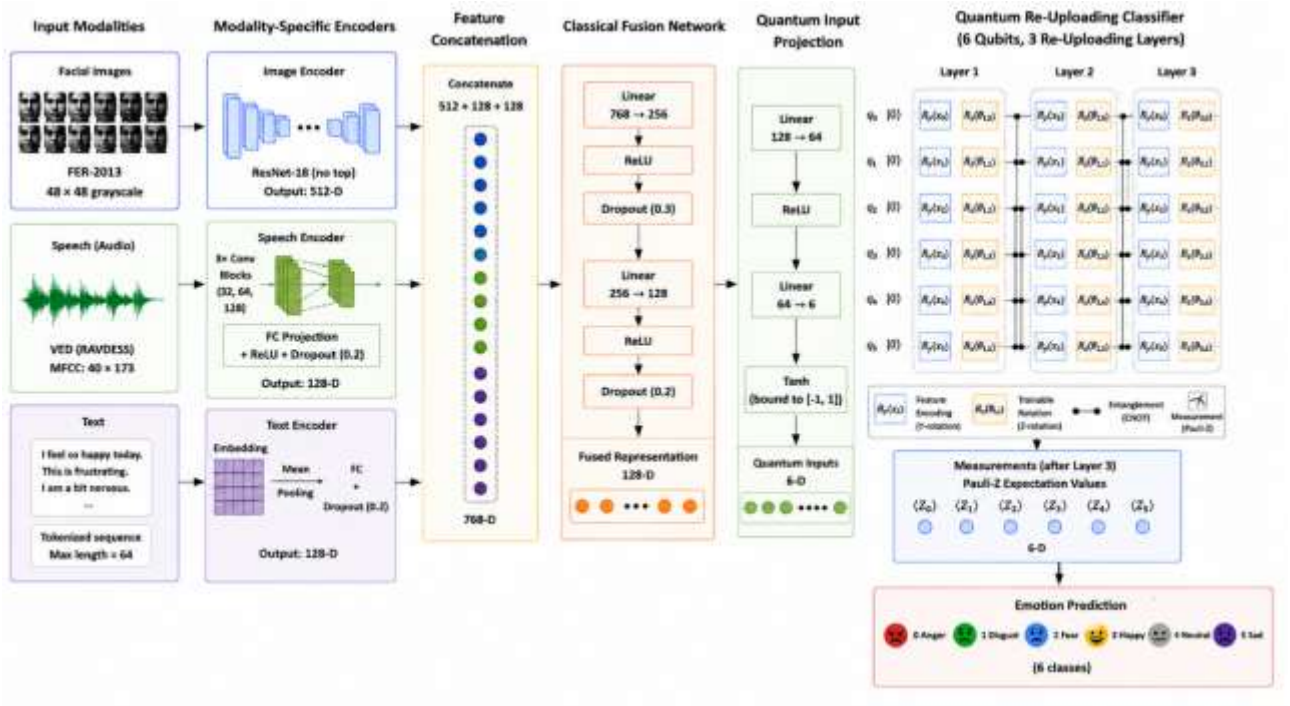

Figure 1. Overall architecture of the proposed hybrid quantum–classical framework for six-class multimodal emotion recognition.

Facial images from FER-2013, speech from VED, and text data are processed by modality-specific encoders to obtain 512-, 128-, and 128-dimensional representations, respectively. The representations are concatenated and transformed through classical fusion to a 128-dimensional representation, which is projected to six quantum inputs and processed by a six-qubit quantum re-uploading circuit with three re-uploading layers for six-class emotion classification.

2.2 Mathematical Formulation of Quantum Re-Uploading

Let h ∈ R^{128} denote the fused multimodal representation produced by the classical fusion network. A trainable quantum input projection maps this representation to a six-dimensional vector x ∈ R^6 , where each component is encoded on one of the six qubits. The projection is implemented as a 128 → 64 → 6 feed-forward network with a final hyperbolic tangent activation, constraining the quantum inputs to the interval [-1,1].

$$x = tanh(W_2\, \sigma\, (W_1 h + b_1) + b_2) \qquad (1)$$

Where:

- h ∈ R^{128} is the fused representation
- W_1 maps 128 → 64
- σ is ReLU
- W_2 maps 64 → 6
- tanh bounds the six quantum inputs

In the proposed quantum re-uploading classifier, the six-dimensional projected feature vector is repeatedly encoded across L = 3 quantum layers. At each layer, the same classical feature vector is re-encoded using data-dependent Y-axis rotations, followed by trainable Z-axis rotations and nearest-neighbor entanglement. Repeated data encoding increases the effective expressivity of the circuit without requiring an increase in the number of qubits [22].

The overall unitary transformation applied to the initial quantum state $|0\rangle^{\otimes d}$ can be expressed as:

$$U(x,\theta) = \prod_{l=1}^{3} U_{ent}\, U_{var}(\theta_l) U_{enc}(x) \qquad (2)$$

Where:

- $U_{enc}(x)$ represents data encoding,
- $U_{var}(\theta_l)$ represents the trainable rotation layer at l,
- U_{ent} represents the nearest-neighbor CNOT entanglement layer.

Feature encoding is performed using Y-axis rotation gates applied to each qubit, with the projected classical features x_i used as the rotation angles.

$$U_{enc}(x) = \otimes_{i=1}^{6} R_y(x_i) \quad (3)$$

where $R_y(x_i)$ represents a rotation around the Y-axis parameterized by the i-th feature.

$$R_y(x_i) = \begin{bmatrix} cos(x_i/2) & -sin(x_i/2) \\ sin(x_i/2) & cos(x_i/2) \end{bmatrix} \quad (4)$$

Trainable quantum operations are introduced through parameterized rotations:

$$U_{var}(\theta_l) = \otimes_{i=1}^{6} R_z(\theta_{l,i}) \quad (5)$$

where $\theta_{l,i}$ is the trainable parameter associated with qubit i in re-uploading layer l. With L = 3 re-uploading layers and six qubits, the circuit contains 3 × 6 = 18 trainable quantum rotation parameters.

Entanglement is introduced using a chain of controlled-NOT (CNOT) gates:

$$U_{ent} = \prod_{i=1}^{5} C\,NOT(i, i+1) \quad (6)$$

A nearest-neighbor CNOT chain is applied after each data-encoding and trainable-rotation stage, establishing interactions among adjacent qubits in a linear connectivity structure.

After the application of all re-uploading layers, measurement is performed in the Pauli-Z basis. The expectation value for each qubit is computed as:

$$z_i = \langle \psi(x,\theta) | Z_i | \psi(x,\theta) \rangle, \quad i = 1, \dots, 6 \quad (7)$$

where the final quantum state is:

$$|\psi(x,\theta)\rangle = U(x,\theta)|0\rangle^{\otimes 6} \quad (8)$$

The Pauli- Z expectation values satisfy $z_i \in [-1,1]$, producing a six-dimensional quantum feature vector z = $[z_1, \dots, z_6]$. The resulting six-dimensional expectation vector z is passed to a classical linear layer that produces six emotion logits. During training, the logits are optimized using cross-entropy loss, while softmax probabilities are used for class interpretation.

2.3 Modality-Specific Encoders

Each modality is processed using a modality-specific encoder designed to extract a compact representation while preserving information relevant to emotion recognition. The image encoder produces a 512-dimensional representation, while the speech and text encoders produce 128-dimensional representations. These modality-specific features are concatenated to form a 768-dimensional multimodal representation, which is subsequently processed by the classical fusion network.

2.3.1 Image Encoder

Facial images are obtained from the FER-2013 dataset and preprocessed as 48×48 grayscale images. The image modality is encoded using a ResNet-18 backbone. The final classification layer is replaced with an identity mapping so that the network outputs a 512-dimensional visual representation after global average pooling. This representation captures discriminative facial features and is subsequently combined with the speech and text representations during multimodal fusion.

2.3.2 Speech Encoder

Speech signals are obtained from the Voice Emotion Dataset (VED), derived from the RAVDESS corpus. Mel-frequency cepstral coefficients (MFCCs) are extracted to represent the acoustic characteristics of the speech signals. The resulting 40 × 173 MFCC representation is processed using a convolutional neural network consisting of three convolutional blocks with 32, 64, and 128 filters, respectively. Each block incorporates batch normalization and ReLU activation, with max-pooling applied after the first two convolutional layers. Adaptive average pooling is applied after the third convolutional layer, followed by a fully connected projection to obtain a 128-dimensional speech representation.

2.3.3 Text Encoder

Textual emotion information is obtained from the Emotion Text Dataset. The text samples are tokenized and converted into integer token sequences with a maximum sequence length of 64. A trainable embedding layer maps the tokens to 128-dimensional representations. The token embeddings are aggregated through mean pooling to obtain a fixed-dimensional textual representation, which is subsequently passed through a projection layer consisting of a 128-to-128 linear transformation, ReLU activation, and dropout. The resulting 128-dimensional vector is used as the textual representation for multimodal fusion.

2.4 Multimodal Fusion Network

The modality-specific representations from the image, speech, and text encoders are concatenated to form a unified multimodal feature vector. The image encoder produces a 512-dimensional representation, while the speech and text encoders each produce 128-dimensional representations. Therefore, the concatenated multimodal representation has a dimensionality of 512 + 128 + 128 = 768.

To integrate the complementary information across modalities, a fully connected fusion network is applied to the 768-dimensional representation. The fusion network consists of two linear layers with ReLU activations and

dropout regularization. The first layer reduces the dimensionality from 768 to 256, followed by ReLU activation and dropout with a rate of 0.3. The second layer further projects the representation from 256 to 128 dimensions, followed by ReLU activation and dropout with a rate of 0.2.

$$h = f_{fusion}\ (x_{img}\ ||x_{speech}\ ||x_{text})$$

where $h \in R^{128}$ represents the fused multimodal representation and || denotes concatenation.
The resulting 128-dimensional fused representation is subsequently passed to a trainable quantum input projection network. This projection maps the 128-dimensional representation through a $128 \rightarrow 64 \rightarrow 6$ feed-forward network, followed by a hyperbolic tangent activation, producing six bounded quantum input features for the six-qubit quantum circuit.

2.5 Quantum Re-Uploading Classifier

The quantum classification stage is implemented using a parameterized variational quantum circuit containing six qubits. The projected six-dimensional feature vector is repeatedly encoded across three quantum re-uploading layers. At each layer, the classical features are encoded using data-dependent Y-axis rotations, followed by trainable Z-axis rotations and nearest-neighbor CNOT entanglement.

Circuit Architecture:

- Number of qubits: 6
- Number of re-uploading layers: 3
- Data encoding: $R_y(x_i)$ on each qubit
- Trainable rotation: $R_z(\theta_{l,i})$
- Trainable quantum parameters: 3 × 6 = 18
- Entanglement: nearest-neighbor CNOT chain with 5 CNOT operations per layer
- Total CNOT applications: 3 × 5 = 15
- Initial State: $|0\rangle^{\otimes 6}$
- Measurement: Pauli- Z expectation value on each qubit, producing six expectation values

In each re-uploading layer, the projected classical features are encoded into the quantum state through Y-axis rotations. Trainable Z-axis rotations then provide adjustable quantum transformations, followed by a nearest-neighbor CNOT chain to introduce interactions between adjacent qubits. Repeating this encoding-and-transformation sequence across three layers increases the effective expressivity of the circuit while maintaining a fixed six-qubit configuration [22].
After the final re-uploading layer, the expectation value of the Pauli-Z operator is measured for each qubit. The resulting six-dimensional expectation vector is passed to a classical linear output layer to produce six emotion-class logits. During training, the model is optimized using cross-entropy loss, while softmax probabilities are used for class interpretation.

2.6 Datasets

Three publicly available benchmark datasets are used to evaluate the proposed multimodal framework across the image, speech, and text modalities. These datasets provide complementary affective information and are summarized in Table 1.

Table 1. Datasets utilized for multimodal emotion recognition

Modality	Dataset	Emotion Classes Used
Image	FER-2013	6
Speech	VED	6
Text	Emotion Text Dataset	6

Each dataset was processed independently according to its modality-specific preprocessing pipeline, and the emotion labels were mapped to a common six-class representation consisting of anger, disgust, fear, happy, neutral, and sad. The resulting samples were organized into consistent training, validation, and test partitions for the multimodal experiment. The same six-class label representation was maintained during classical fusion and subsequent quantum classification.

Table 2. Multimodal dataset splits used in the experiments

Training	Validation	Test	Total
20,407	5,131	6,347	31,885

2.7 Preprocessing and Label Alignment

Each modality is preprocessed according to its specific data characteristics before feature extraction. Facial images from FER-2013 are converted to 48 × 48 grayscale representations and normalized before being provided to the image encoder. Speech samples from the Voice Emotion Dataset (VED) are represented using 40-dimensional Mel-frequency cepstral coefficients (MFCCs), with a fixed temporal dimension of 173 frames. Text samples from the Emotion Text Dataset are converted into token sequences with a maximum sequence length of 64 and processed using a trainable embedding representation followed by mean pooling and projection. These preprocessing steps produce fixed-dimensional modality-specific representations suitable for multimodal fusion.

Since the datasets originate from different sources and may use different label conventions, their emotion labels are mapped to a common six-class representation before multimodal fusion. The six emotion categories used in the final experiments are anger, disgust, fear, happy, neutral, and sad. Samples with incompatible or unavailable labels are excluded during dataset construction. This label mapping ensures that the same emotion-class definition is maintained across the image, speech, and text modalities.

The resulting modality-specific samples are organized into training, validation, and test partitions. The same six-class label space is maintained throughout the classical multimodal model and the quantum re-uploading classifier, enabling direct comparison between the classical and quantum approaches.

2.8 Experimental Setup

2.8.1 Model Training and Selection

The classical multimodal models and the quantum re-uploading classifier were trained separately using the training partition, while the validation partition was used for model selection and monitoring generalization performance. The test partition was kept completely separate and was used only for the final evaluation of the selected models.

The classical multimodal models were trained for 10 epochs using a batch size of 32 with the AdamW optimizer, a learning rate of 0.0001, and a weight decay of 0.0001. For each epoch, performance was evaluated on the validation set, and the checkpoint with the highest validation accuracy was retained for subsequent test evaluation. The same training and validation procedure was applied to the different classical modality configurations used in the ablation study.

The quantum re-uploading classifier was also trained for 10 epochs with a batch size of 32 using the Adam optimizer with a learning rate of 0.001. The model was evaluated on the validation set after each epoch, and the checkpoint achieving the highest validation accuracy was selected for final testing. The selected checkpoint was then evaluated once on the held-out test partition.

2.8.2 Model Configuration

The final multimodal framework uses modality-specific encoders for facial images, speech, and text. The image encoder produces a 512-dimensional visual representation, while the speech and text encoders each produce 128-dimensional representations. These modality-specific representations are concatenated to form a 768-dimensional multimodal feature vector.

The concatenated representation is processed by a classical fusion network consisting of fully connected layers with ReLU activation and dropout regularization. The fusion network reduces the 768-dimensional representation to a 128-dimensional fused representation. This representation is subsequently passed through a quantum input projection network that maps the features from 128 dimensions to 64 dimensions and finally to 6 dimensions. The resulting six-dimensional vector serves as the input to the six-qubit quantum re-uploading classifier.

This configuration allows the heterogeneous image, speech, and text representations to be integrated before quantum processing while maintaining a compact feature representation for the quantum circuit.

2.8.3 Quantum Circuit Configuration

The quantum classifier is implemented as a parameterized variational quantum circuit with six qubits. Each qubit receives one component of the projected six-dimensional multimodal feature vector. A quantum re-uploading strategy is adopted with three re-uploading layers, in which the classical features are repeatedly encoded into the circuit using data-dependent Y-axis rotations.

Each re-uploading layer consists of feature-dependent Y-axis rotation gates followed by trainable Z-axis rotations and nearest-neighbor entanglement implemented using a chain of CNOT gates. With six qubits and three re-uploading layers, the circuit contains 18 trainable quantum rotation parameters and applies five nearest-neighbor CNOT operations per layer, resulting in 15 CNOT applications across the three layers.

After the final re-uploading layer, measurement is performed in the Pauli-Z basis. The expectation value of the Pauli-Z operator is obtained for each qubit, producing a six-dimensional quantum feature vector. These expectation values are passed to a classical linear output layer to generate the six emotion-class logits. During training, the logits are optimized using cross-entropy loss, while softmax probabilities are used for class interpretation.

2.8.4 Training Protocol

The classical multimodal models were trained using class-weighted cross-entropy loss, with class weights derived from the six-class FER-2013 training distribution. The quantum re-uploading classifier was trained using standard

cross-entropy loss. The classical models were optimized using AdamW with a learning rate of 0.0001 and weight decay of 0.0001, whereas the quantum re-uploading classifier was optimized using Adam with a learning rate of 0.001. All models were trained for 10 epochs, with validation performance monitored after each epoch. The checkpoint with the highest validation accuracy was retained for final evaluation on the held-out test set.

For the classical ablation experiments, the same 10-epoch training protocol was applied to the four modality configurations: Image Only, Image + Text, Image + Speech, and Image + Speech + Text. The best validation checkpoint was selected independently for each configuration before test evaluation.

The quantum re-uploading classifier was trained using the same batch size and 10-epoch schedule. Its validation performance was monitored after each epoch, and the checkpoint achieving the highest validation accuracy was selected for final testing.

Table 3. Training configuration

Parameter	Value
Classical loss function	Class-weighted cross-entropy
Quantum loss function	Cross-entropy
Batch size	32
Training epochs	10
Classical speech dropout	0.2
Classical text dropout	0.2
Fusion dropout after 768 → 256	0.3
Fusion dropout after 256 → 128	0.2
Quantum backend	State-vector simulator
Classical optimizer	AdamW
Classical learning rate	0.0001
Classical weight decay	0.0001
Quantum optimizer	Adam
Quantum learning rate	0.001

2.8.5 Evaluation Metrics

Classification accuracy and macro-averaged F1-score are used as the primary evaluation metrics. Accuracy measures the proportion of correctly classified samples over the complete test set. The macro-averaged F1-score is calculated by computing the F1-score independently for each emotion class and then taking their unweighted mean, giving equal importance to all classes.

To further analyze model behavior across emotion categories, confusion matrices and per-class precision, recall, and F1-scores are also examined. These analyses provide additional insight into class-specific recognition performance and the emotion categories that are most frequently confused.

2.8.6 Reproducibility Considerations

The experimental configuration, dataset partitions, model architectures, training settings, and evaluation procedure are documented to support reproducibility. The classical and quantum models are trained using the fixed training and validation partitions described in the preceding sections, while the held-out test partition is used only for final evaluation.

The quantum circuit is simulated using a state-vector quantum backend, allowing the proposed circuit to be evaluated without requiring physical quantum hardware. The final reported results correspond to the selected checkpoints obtained from the validation-based model selection procedure.

3. Results and Discussion

3.1 Classical Model Training and Convergence

The training behavior of the proposed Image + Speech + Text multimodal model was analyzed over 10 epochs using training loss, training accuracy, and validation accuracy. As shown in Figure 2, the training loss decreases consistently throughout the training process, from 1.4853 in the first epoch to 0.1570 in the tenth epoch. The continuous reduction in loss indicates effective optimization of the multimodal classification network using the AdamW optimizer with a learning rate of 0.0001.

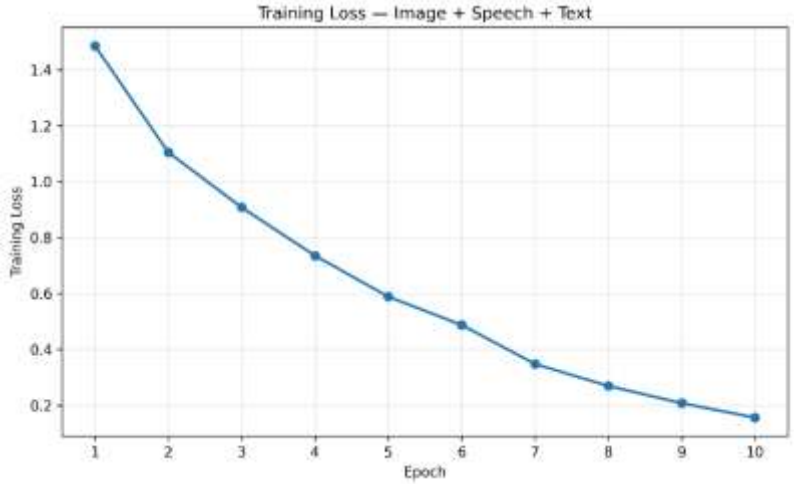

Figure 2. Training loss of the Image + Speech + Text multimodal model over 10 epochs.

The corresponding training and validation accuracy curves are presented in Figure 3. Training accuracy increased steadily from 41.69% in the first epoch to 94.19% by epoch 10. Validation accuracy also improved during training, increasing from 56.29% in the first epoch to a maximum of 72.62% at epoch 9. The validation accuracy was 71.22% at epoch 10, indicating a slight fluctuation after reaching its best performance.

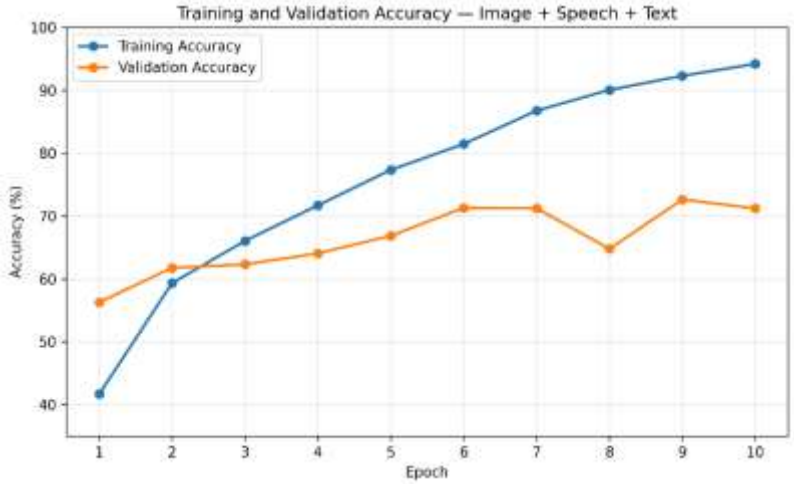

Figure 3. Training and validation accuracy of the Image + Speech + Text multimodal model over 10 epochs.

The best-performing checkpoint was therefore selected based on validation accuracy at epoch 9, rather than simply using the final epoch. The difference between training and validation accuracy becomes more pronounced during the later epochs, particularly after epoch 6, suggesting increasing specialization of the model to the training data. However, the validation accuracy remains above 70% from epochs 6–10 except for epoch 8, indicating that the model retains reasonable generalization performance despite the increasing training–validation gap.
Overall, the learning curves demonstrate that the classical multimodal model converges effectively within the 10-epoch training schedule. The validation-based checkpoint selection provides a consistent basis for subsequent evaluation on the held-out test set.

3.2 Classical Ablation and Multimodal Performance

To evaluate the contribution of the individual modalities, an ablation study was conducted using four classical configurations: Image Only, Image + Text, Image + Speech, and Image + Speech + Text. All configurations were trained using the same training and validation partitions and evaluated on the same held-out test set of 6,347 samples. The resulting test performance is summarized in Table 4 and illustrated in Figure 4.
The Image Only configuration achieved a test accuracy of 48.31% and a macro F1-score of 47.16%, establishing the unimodal visual baseline. Incorporating text alongside image features increased the test accuracy to 60.82% and the macro F1-score to 58.11%. Similarly, combining image and speech features resulted in 61.29% accuracy

and a macro F1-score of 56.54%. These results demonstrate that the addition of either textual or speech information provides a substantial improvement over the image-only configuration.

Table 4. Classical ablation performance across different modality configurations on the held-out test set.

Model	Test Accuracy (%)	Macro F1-Score (%)
Image Only	48.31	47.16
Image + Text	60.82	58.11
Image + Speech	61.29	56.54
Image + Speech + Text	73.94	70.64

The complete Image + Speech + Text configuration achieved the highest performance, with a test accuracy of 73.94% and a macro F1-score of 70.64%. This represents an improvement of 25.63 percentage points in accuracy over the Image Only baseline. It also exceeds the Image + Text and Image + Speech configurations by 13.12 and 12.65 percentage points, respectively. The macro F1-score also increased from 47.16% for the Image Only configuration to 70.64% for the complete multimodal model, representing an improvement of 23.48 percentage points and indicating that the performance gain is not limited to overall accuracy but also reflects improved classification across the emotion categories.

The progressive improvement obtained by incorporating multiple modalities indicates that speech and textual information provide complementary affective cues to the visual representation. The substantially higher performance of the three-modality configuration suggests that combining these heterogeneous representations provides complementary information that can be exploited by the classical fusion network, which is not available from a single modality or a two-modality combination.

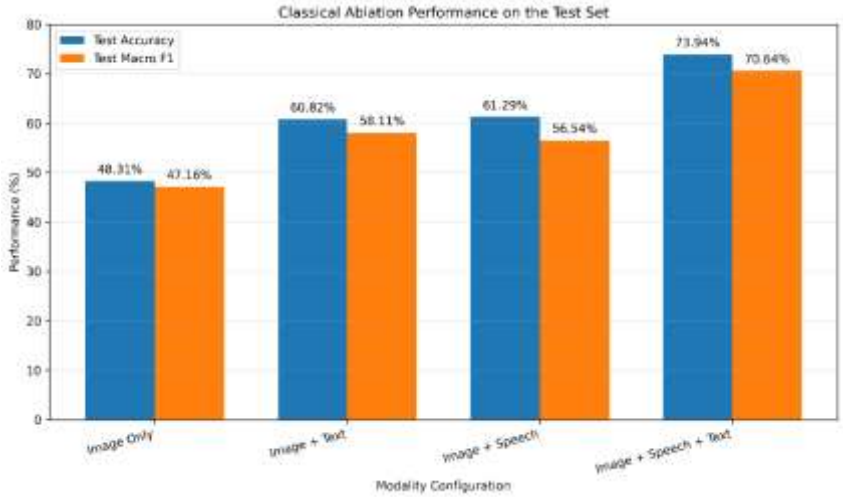

Figure 4. Classical ablation performance across different modality configurations on the held-out test set.

3.3 Per-Class Performance and Confusion Analysis

The class-wise performance of the complete Image + Speech + Text model was analyzed using precision, recall, F1-score, and class-wise accuracy on the held-out test set. The model achieved an overall test accuracy of 73.94% and a macro F1-score of 70.64%. As shown in Table 5, the performance varies across the six emotion categories, indicating that some emotions are more consistently recognized than others.

Table 5. Per-class performance of the classical multimodal model.

Emotion	Precision (%)	Recall (%)	F1-Score (%)	Class Accuracy (%)
Anger	58.13	57.83	57.98	57.83
Disgust	70.37	51.35	59.38	51.35
Fear	64.18	73.14	68.37	73.14
Happy	80.45	78.18	79.30	78.18
Neutral	83.83	82.81	83.31	82.81
Sad	76.83	74.18	75.48	74.18
Macro Average	72.30	69.58	70.64	--

Neutral was the strongest-performing emotion, achieving a class accuracy of 82.81% and an F1-score of 83.31%. Happy also showed strong recognition performance, with an accuracy of 78.18% and an F1-score of 79.30%. Sad

and Fear achieved accuracies of 74.18% and 73.14%, respectively, indicating relatively consistent recognition of these categories. In contrast, Disgust obtained the lowest class accuracy at 51.35%, followed by Anger at 57.83%. The lower recall for Disgust (51.35%) indicates that a considerable proportion of disgust samples were assigned to other emotion categories.

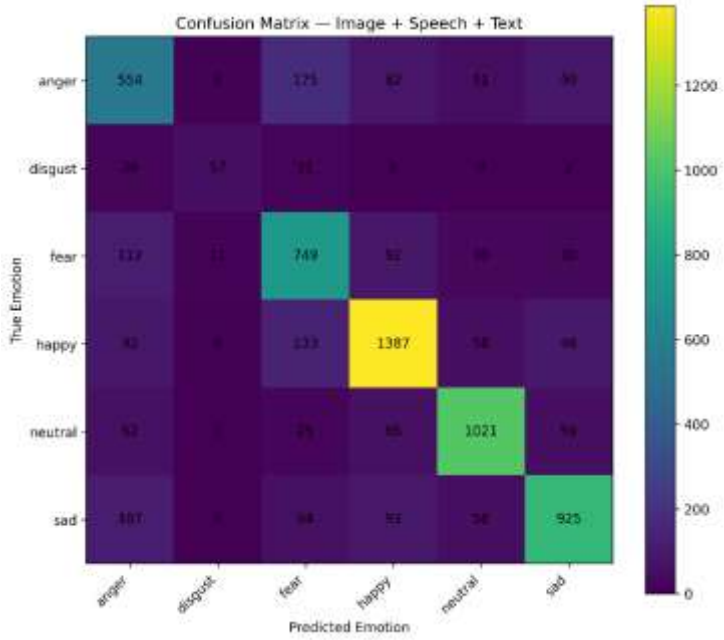

Figure 5. Confusion matrix of the Image + Speech + Text classical multimodal model on the held-out test set.

The confusion matrix in Figure 5 provides further insight into the class-wise prediction behavior of the multimodal model. The diagonal entries indicate correctly classified samples, while the off-diagonal entries represent misclassifications between emotion categories. Strong diagonal values are observed for Happy, Neutral, and Sad, consistent with their relatively high class-wise accuracies.

The largest confusion involving Anger occurs with Fear, with 175 anger samples classified as Fear. Conversely, 112 Fear samples were classified as Anger. This represents approximately 18.3% of the anger samples and 10.9% of the fear samples, respectively. The two-way confusion indicates that these categories remain comparatively difficult to separate using the learned multimodal representations.

Fear and Happy also exhibit some cross-class confusion, with 92 Fear samples classified as Happy and 133 Happy samples classified as Fear. Similarly, 93 Sad samples were classified as Happy, while 65 Neutral samples were classified as Happy. These off-diagonal predictions indicate that the model occasionally confuses emotionally related or less distinctly represented samples despite the use of multiple modalities.

Disgust presents a different pattern. Although its precision is relatively high at 70.37%, its recall is only 51.35%, indicating that the model is conservative when predicting this class. Of the 111 disgust samples, 57 were correctly classified, while 26 were predicted as Anger and 21 as Fear. This accounts for much of the reduced recall observed for the Disgust category.

Overall, the confusion matrix and class-wise metrics show that multimodal fusion substantially improves recognition across the six emotion categories, but performance remains dependent on the discriminative characteristics of individual emotions. The stronger results for Neutral, Happy, and Sad contrast with the lower recognition of Disgust and Anger. These findings demonstrate that multimodal information can provide complementary cues while not completely eliminating ambiguity between emotion categories.

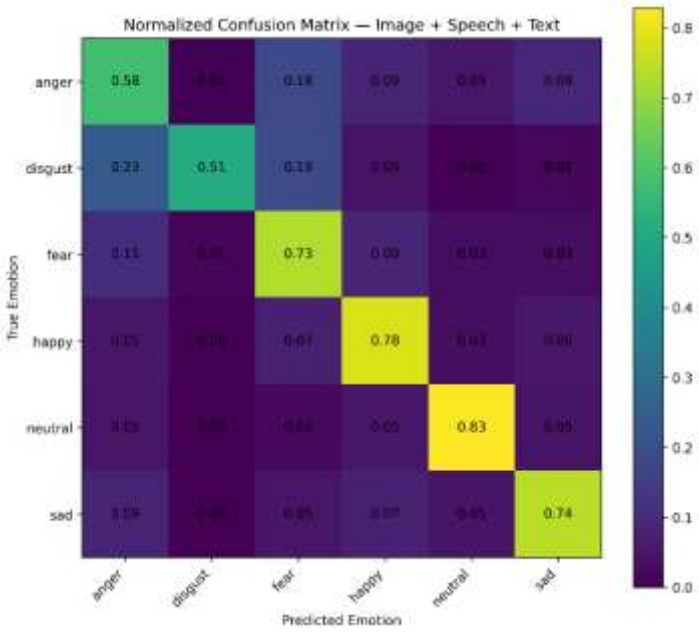

Figure 6. Normalized confusion matrix of the Image + Speech + Text classical multimodal model on the held-out test set.

The normalized confusion matrix in Figure 6 provides a class-wise view of the prediction distribution by converting the raw counts into proportions within each true emotion category. The diagonal values correspond to the recall of each emotion and are consistent with the class-wise accuracies reported in Table 5. Neutral achieves the highest normalized diagonal value (0.83), followed by Happy (0.78), Sad (0.74), and Fear (0.73). Anger and Disgust show lower diagonal values of 0.58 and 0.51, respectively, confirming that these classes are more difficult for the model to distinguish.

The normalized matrix also highlights the main sources of class confusion. For Anger, 18% of the samples are classified as Fear, while 11% of Fear samples are classified as Anger. Disgust shows particularly high confusion with Anger and Fear, with 23% and 19% of its samples assigned to these categories, respectively. These normalized proportions are consistent with the raw confusion counts and further demonstrate that the lower performance of Anger and Disgust is primarily associated with their confusion with other negative or closely related emotion categories.

Overall, the class-wise results demonstrate that the multimodal model provides strong recognition for Neutral, Happy, Sad, and Fear, while Anger and Disgust remain more challenging. The combination of raw and normalized confusion matrices provides complementary evidence of both the magnitude and distribution of classification errors across the six emotion categories.

3.4 Quantum Re-Uploading Classifier Training and Performance

The training behavior of the quantum re-uploading classifier was analyzed over 10 epochs using training and validation accuracy. As shown in Figure 7, the training accuracy increased rapidly from 75.99% in the first epoch to 84.68% in the second epoch, after which it remained relatively stable around 86%. This indicates that the classifier learned the projected multimodal representation rapidly during the initial training stages.

The validation accuracy remained comparatively stable throughout the training process, ranging from 70.00% to 71.55%. The highest validation accuracy of 71.55%, together with a validation macro F1-score of 67.10%, was obtained at epoch 9. Although the training accuracy continued to remain around 86%, the validation accuracy decreased to 70.91% at epoch 10. Therefore, the epoch-9 checkpoint was selected based on validation accuracy rather than simply using the final training epoch.

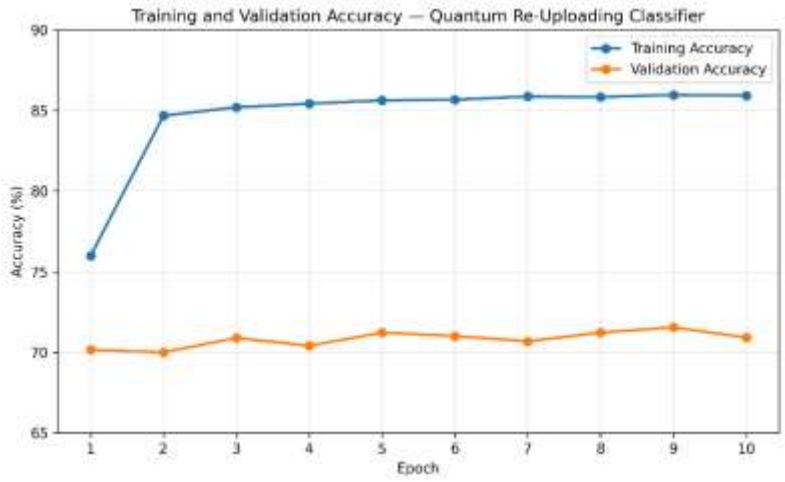

Figure 7. Training and validation accuracy of the quantum re-uploading classifier over 10 epochs.

The selected checkpoint was subsequently evaluated on the held-out test set containing 6,347 samples. The quantum re-uploading classifier achieved a test accuracy of 72.57% and a macro F1-score of 68.76%. The selected checkpoint achieved a test accuracy of 72.57%, slightly higher than its best validation accuracy of 71.55%, indicating that the selected model maintained good generalization on the held-out test set.

The results also show that the quantum classifier reaches a stable training regime within the adopted 10-epoch schedule. The validation performance does not increase monotonically, but remains within a narrow range after the initial epochs, with the best performance occurring at epoch 9. This supports the use of validation-based checkpoint selection for the final quantum evaluation.

3.5 Quantum Per-Class Performance and Confusion Analysis

The class-wise performance of the quantum re-uploading classifier was further analyzed using precision, recall, F1-score, and class-wise accuracy on the held-out test set. The model achieved an overall test accuracy of 72.57% and a macro F1-score of 68.76%. As shown in Table 6, the performance varies across the six emotion categories, indicating that the quantum classifier recognizes some emotional categories more consistently than others.

Table 6. Per-class performance of the quantum re-uploading classifier.

Emotion	Precision (%)	Recall (%)	F1-Score (%)	Class Accuracy (%)
Anger	52.81	50.00	51.37	50.00
Disgust	73.91	45.95	56.67	45.95
Fear	66.48	70.31	68.34	70.31
Happy	76.16	79.43	77.76	79.43
Neutral	84.44	83.21	83.82	83.21
Sad	75.31	73.86	74.57	73.86
Macro Average	71.52	67.13	68.76	--

The Neutral category achieved the strongest overall class-wise performance, with a precision of 84.44%, recall of 83.21%, and F1-score of 83.82%. The Happy category also showed strong recognition performance, achieving a recall of 79.43% and an F1-score of 77.76%. The relatively high performance of these categories indicates that their multimodal representations provide more distinguishable patterns for the classifier. In contrast, Disgust and Anger were the most challenging categories. Disgust achieved a recall of only 45.95%, while Anger achieved a recall of 50.00%. The lower recall values indicate that a substantial proportion of samples from these categories were assigned to other emotion classes.

Figure 8. Confusion matrix of the quantum re-uploading classifier on the held-out test set.

The confusion matrix in Figure 8 provides further insight into the prediction behavior of the quantum classifier. The diagonal entries represent correctly classified samples, whereas the off-diagonal entries represent misclassifications. Strong diagonal values are observed for Happy, Neutral, and Sad, consistent with their relatively high class-wise recall values reported in Table 6.

The largest confusion involving Anger occurs with Fear, with 164 Anger samples classified as Fear. Anger is also confused with Happy and Sad, with 149 and 120 samples, respectively. Conversely, 129 Fear samples were classified as Anger. This two-way confusion indicates that the learned multimodal representations do not always provide sufficiently discriminative information to separate these categories.

Confusion between Fear and Happy is also observed, with 129 Fear samples classified as Happy and 120 Happy samples classified as Fear. Although the two categories differ semantically, their observed facial and speech characteristics can overlap in high-arousal expressions, making them more difficult to distinguish from the available representations.

The Sad category shows relatively strong recognition, with 921 correctly classified samples, but notable errors occur toward Anger and Happy, with 123 and 87 samples, respectively. Similarly, Disgust has only 51 correctly classified samples, with 29 samples predicted as Anger and 18 as Fear. This explains its relatively low recall of 45.95%, despite its comparatively high precision of 73.91%.

Figure 9. Normalized confusion matrix of the quantum re-uploading classifier on the held-out test set.

The normalized confusion matrix in Figure 9 presents the same prediction behavior in proportional form, allowing class-wise comparisons independent of the different numbers of test samples in each emotion category. The diagonal values correspond directly to the recall of each class. The highest diagonal values are observed for Neutral (0.83) and Happy (0.79), followed by Sad (0.74) and Fear (0.70). In contrast, Anger (0.50) and Disgust (0.46) show substantially lower diagonal values, confirming that these categories are more difficult for the quantum classifier to recognize correctly.

Overall, the class-wise results demonstrate that the quantum re-uploading classifier does not perform uniformly across all emotion categories. The stronger recognition of Neutral, Happy, and Sad contrasts with the lower recognition of Anger and Disgust. The confusion patterns indicate that errors are primarily associated with overlap between emotion categories rather than a single dominant failure mode. These findings complement the overall test accuracy and macro F1-score by showing how the classifier behaves at the individual emotion level.

3.6 Classical–Quantum Performance Comparison

The performance of the classical multimodal model and the proposed quantum re-uploading classifier was compared using the held-out test set. The classical ablation experiments demonstrated that incorporating complementary modalities progressively improved classification performance. The Image Only configuration achieved a test accuracy of 48.31% and a macro F1-score of 47.16%, while the Image + Text and Image + Speech configurations achieved 60.82% and 61.29% accuracy, respectively. The complete Image + Speech + Text configuration produced the highest classical performance, achieving a test accuracy of 73.94% and a macro F1-score of 70.64%.

The quantum re-uploading classifier achieved a test accuracy of 72.57% and a macro F1-score of 68.76% on the same held-out test set. Thus, the quantum classifier remained close to the best-performing classical multimodal configuration, with an accuracy difference of only 1.37 percentage points and a macro F1 difference of 1.88 percentage points. This relatively small gap indicates that the proposed quantum re-uploading approach was able to retain a substantial portion of the classification performance achieved by the classical multimodal architecture.

The comparison also highlights the importance of multimodal information. The classical Image Only baseline achieved 48.31% accuracy, whereas combining image, speech, and text increased the accuracy to 73.94%, representing an improvement of 25.63 percentage points. The quantum classifier, which operates on the projected multimodal representation, achieved 72.57%, further demonstrating that the combined representation provides informative features for quantum-based classification.

The Macro F1 results show a similar pattern. The complete classical multimodal model achieved 70.64%, compared with 68.76% for the quantum re-uploading classifier. The relatively small difference between the two models suggests that the quantum classifier does not merely achieve comparable overall accuracy through a biased prediction toward frequently occurring classes; its macro-averaged performance also remains close to that of the classical model.

Table 7. Parameter comparison between the classical and quantum models.

Model	Trainable Parameters	Parameters (M)
Classical Image + Speech + Text	15,440,006	15.4400
Quantum Re-Uploading	8,706	0.0087

The parameter count further highlights the compactness of the quantum re-uploading model. The classical Image + Speech + Text model contains 15,440,006 trainable parameters, whereas the quantum re-uploading model contains only 8,706 trainable parameters. Thus, the quantum model uses approximately 99.94% fewer trainable parameters, corresponding to approximately 1,773 times as many trainable parameters in the classical model. Despite this substantial reduction in parameter count, the quantum model achieved a test accuracy of 72.57% and a macro F1-score of 68.76%, compared with 73.94% and 70.64%, respectively, for the classical model. These results indicate that the quantum re-uploading classifier provides competitive predictive performance with a substantially more compact trainable model.

Overall, the results demonstrate that the quantum re-uploading classifier provides competitive performance relative to the classical multimodal baseline under the current experimental configuration. The classical model remains slightly superior in both accuracy and macro F1-score, but the gap is considerably smaller than what was reported in the earlier manuscript. Therefore, the current results support positioning quantum re-uploading as a competitive alternative for multimodal emotion classification, rather than claiming that it outperforms the classical approach.

3.7 Discussion and Implications

The experimental results demonstrate that multimodal integration substantially improves emotion recognition performance compared with using a single modality. In the classical ablation study, the Image Only configuration achieved 48.31% test accuracy, whereas the combination of Image + Speech + Text achieved 73.94%. The

improvement indicates that speech and textual information provide complementary affective cues that enhance the visual representation. The intermediate performance of the Image + Text and Image + Speech configurations further support the contribution of additional modalities to the final classification performance.

The proposed quantum re-uploading classifier achieved 72.57% test accuracy and a macro F1-score of 68.76%, compared with 73.94% accuracy and 70.64% macro F1-score for the classical multimodal model. Although the classical model achieved slightly higher performance, the difference was relatively small, at 1.37 percentage points in accuracy and 1.88 percentage points in macro F1-score. This result indicates that the quantum re-uploading classifier was able to utilize the projected multimodal representation effectively and achieve performance close to that of the corresponding classical fusion model.

The class-wise results provide additional insight into the behavior of the quantum classifier. Neutral achieved the highest F1-score of 83.82%, while Happy achieved an F1-score of 77.76%. These results indicate that the model can identify emotion categories with relatively distinctive multimodal patterns more reliably. In contrast, Anger and Disgust were more challenging, with F1-scores of 51.37% and 56.67%, respectively. The confusion matrices show that these lower scores are associated with substantial misclassification toward other negative or affectively related categories. For example, 164 Anger samples were classified as Fear, while 129 Fear samples were classified as Anger.

The comparison between the classical and quantum models should therefore be interpreted in terms of both overall performance and class-wise behavior. The classical model provides the strongest overall result in the present experiments, while the quantum re-uploading model achieves a closely comparable level of performance. The quantum model should consequently not be presented as outperforming the classical architecture; instead, the results support its potential as a competitive quantum-based classification approach for multimodal emotion recognition.

The observed performance also highlights the importance of the multimodal feature representation supplied to the classifier. The substantial improvement from the Image Only baseline to the complete Image + Speech + Text configuration suggests that heterogeneous modalities contain complementary information that cannot be captured as effectively by an individual modality. The quantum classifier operates on this projected multimodal representation, indicating that quantum processing can be incorporated after classical modality-specific feature extraction rather than requiring the entire multimodal pipeline to be implemented quantum mechanically.

Finally, the present experiments were conducted using a state-vector quantum simulator. Consequently, the reported quantum results should be interpreted as simulation-based results rather than direct evidence of performance on physical quantum hardware. Real quantum devices introduce noise, decoherence, and sampling effects that may affect classification performance. Evaluation on NISQ hardware, investigation of noise-aware training, and exploration of alternative quantum encoding and circuit designs are therefore important directions for future work.

4. Conclusions

This study presented a hybrid quantum-classical framework for multimodal emotion recognition using image, speech, and text modalities. Modality-specific feature representations were combined through a classical fusion pipeline and subsequently projected into a six-dimensional representation for quantum processing. A quantum re-uploading classifier was then employed to repeatedly encode the projected multimodal features within a variational quantum circuit.

The classical ablation experiments demonstrated the benefit of multimodal integration. The Image Only configuration achieved a test accuracy of 48.31% and a macro F1-score of 47.16%, while the complete Image + Speech + Text configuration achieved 73.94% test accuracy and 70.64% macro F1-score. The 25.63-percentage-point improvement in accuracy demonstrates the benefit of combining complementary information from multiple modalities.

The proposed quantum re-uploading classifier achieved a test accuracy of 72.57% and a macro F1-score of 68.76% on the same held-out test set. The best validation performance was obtained at epoch 9, with a validation accuracy of 71.55% and a macro F1-score of 67.10%. The quantum model therefore remained close to the classical multimodal model, with differences of only 1.37 percentage points in accuracy and 1.88 percentage points in macro F1-score. Although the classical model achieved the higher absolute performance, the results demonstrate that the quantum re-uploading classifier can effectively process the projected multimodal representation and provide competitive classification performance.

The class-wise analysis further showed that emotion recognition performance varies across categories. For the quantum classifier, Neutral achieved the highest F1-score of 83.82%, followed by Happy at 77.76%, whereas Anger and Disgust were more challenging, with F1-scores of 51.37% and 56.67%, respectively. The confusion-matrix analysis showed that these lower-performing categories were affected by substantial cross-class confusion, highlighting the difficulty of distinguishing emotionally overlapping patterns.

The quantum experiments were performed using a quantum simulator; therefore, the reported results represent simulation-based performance rather than evaluation on physical quantum hardware. Future work will focus on

validating the proposed approach on NISQ devices, investigating noise-aware training and alternative quantum feature-encoding strategies, and exploring improved multimodal alignment and fusion mechanisms.

5. Acknowledgements
The authors would like to thank the providers of the publicly available datasets used in this research, including the VED for speech data, the Emotion Text Dataset for textual data, and the FER-2013 dataset for facial images. The authors also acknowledge the open-source quantum machine learning libraries PennyLane and PyTorch that supported this research.

Author Contributions: Conceptualization, S.S.N.C.P. and S.S.H.; methodology, S.S.N.C.P.; software, S.S.N.C.P.; validation, S.S.N.C.P. and S.S.H.; formal analysis, S.S.N.C.P.; investigation, S.S.N.C.P.; resources, S.S.H.; data curation, S.S.N.C.P.; writing—original draft preparation, S.S.N.C.P.; writing—review and editing, S.S.H.; visualization, S.S.N.C.P.; supervision, S.S.H.; project administration, S.S.H.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding acquisition: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Rajapakshe, T.; Rana, R.; Riaz, F.; Khalifa, S.; Schuller, B. W. Representation Learning with Parameterised Quantum Circuits for Advancing Speech Emotion Recognition. Sci. Rep. 2025, 15 (1). DOI: 10.1038/s41598-025-27871-4.
[2] Norval, M.; Wang, Z. Quantum AI in Speech Emotion Recognition. Entropy 2025, 27 (12), 1201. DOI: 10.3390/E27121201.
[3] Alsubai, S.; Alqahtani, A.; Alanazi, A.; Sha, M.; Gumaei, A. Facial Emotion Recognition Using Deep Quantum and Advanced Transfer Learning Mechanism. Front. Comput. Neurosci. 2024, 18, 1435956. DOI: 10.3389/FNCOM.2024.1435956.
[4] Wang, Z.; Wang, Y. Emotion Recognition Based on Multimodal Physiological Electrical Signals. Front. Neurosci. 2025, 19, 1512799. DOI: 10.3389/FNINS.2025.1512799.
[5] Sun, W.; Yan, X.; Su, Y.; Wang, G.; Zhang, Y. MSDSANet: Multimodal Emotion Recognition Based on Multi-Stream Network and Dual-Scale Attention Network Feature Representation. Sensors 2025, 25 (7), 2029. DOI: 10.3390/S25072029.
[6] Wu, Y.; Zhang, S.; Li, P. Multi-Modal Emotion Recognition in Conversation Based on Prompt Learning with Text-Audio Fusion Features. Sci. Rep. 2025, 15 (1), 8855. DOI: 10.1038/S41598-025-89758-8.
[7] Saif, M.; Onim, H.; Humble, T. S.; Thapliyal, H. Emotion Recognition in Older Adults with Quantum Machine Learning and Wearable Sensors. arXiv 2025. DOI: 10.48550/arXiv.2507.08175.
[8] Dutta, S.; Balaji, S.; Ganapathy, S. A Mixture-of-Experts Model for Multimodal Emotion Recognition in Conversations. arXiv 2026. DOI: 10.48550/arXiv.2602.23300.
[9] Han, J.; et al. Pioneering Multimodal Emotion Recognition in the Era of Large Models: From Closed Sets to Open Vocabularies. arXiv 2025. DOI: 10.48550/arXiv.2512.20938.
[10] Qu, C.; et al. Enhancing Emotion Recognition in Virtual Reality: A Multimodal Dataset and a Temporal Emotion Detector. Front. Psychol. 2025, 16, 1709943. DOI: 10.3389/FPSYG.2025.1709943.
[11] Zhang, Y.; Tao, X.; Ai, H.; Chen, T.; Gan, Y. Multimodal Emotion Recognition by Fusing Video Semantic in MOOC Learning Scenarios. Lecture Notes in Computer Science 2024, 15289, 1–15. DOI: 10.1007/978-981-96-6585-3_1.
[12] Lian, Z.; et al. MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition. In Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing; 2024, 41–48. DOI: 10.1145/3689092.3689959.
[13] Waligora, P.; et al. Joint Multimodal Transformer for Emotion Recognition in the Wild. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2024, 4625–4635. DOI: 10.1109/CVPRW63382.2024.00465.
[14] Wu, C.; et al. Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects. Findings of the Association for Computational Linguistics: EMNLP 2025, 6257–6274. DOI: 10.18653/v1/2025.findings-emnlp.332.
[15] Gül, B. C.; Nadig, S.; Tziampazis, S.; Jazdi, N.; Weyrich, M. FedMultiEmo: Real-Time Emotion Recognition via Multimodal Federated Learning. arXiv 2025. DOI: 10.48550/arXiv.2507.15470.
[16] Liu, F. Artificial Intelligence in Emotion Quantification: A Prospective Overview. CAAI Artif. Intell. Res. 2024, 3, 9150040. DOI: 10.26599/AIR.2024.9150040.

[17] Yazici, A.; Kucukyilmaz, T.; Dokeroglu, T.; Sharipbay, A.; Lee, M. H.; Tyler, B. State-of-the-Art Multimodal Emotion Recognition: A Comprehensive Survey and Taxonomy. Intell. Syst. Appl. 2026, 30, 200642. DOI: 10.1016/J.ISWA.2026.200642.
[18] Ramaswamy, M. P. A.; Palaniswamy, S. Multimodal Emotion Recognition: A Comprehensive Review, Trends, and Challenges. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2024, 14 (6). DOI: 10.1002/WIDM.1563.
[19] Chandanwala, A. A.; Bhowmik, S.; Chaudhury, P.; Pravin, S. C. Hybrid Quantum Deep Learning Model for Emotion Detection Using Raw EEG Signal Analysis. arXiv 2024. DOI: 10.48550/arXiv.2411.17715.
[20] Pokharel, A.; Rahman, R.; Morris, T.; Nguyen, D. C. Quantum Federated Learning for Multimodal Data: A Modality-Agnostic Approach. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2025, 545–554. DOI: 10.1109/CVPRW67362.2025.00059.
[21] Florestiyanto, M. Y.; Surjono, H. D.; Jati, H. Benchmarking Quantum Kernels and Modern Vision Models for Compound Facial Expression Recognition. Sci. Rep. 2026. DOI: 10.1038/s41598-026-41514-2.
[22] Pérez-Salinas, A.; Cervera-Lierta, A.; Gil-Fuster, E.; Latorre, J. I. Data Re-Uploading for a Universal Quantum Classifier. Quantum 2020, 4, 226. DOI: 10.22331/q-2020-02-06-226.
[23] Dave, K.; Innan, N.; Behera, B. K.; Mumtaz, Z.; Al-Kuwari, S.; Farouk, A. SentiQNF: A Novel Approach to Sentiment Analysis Using Quantum Algorithms and Neuro-Fuzzy Systems. IEEE Trans. Comput. Soc. Syst. 2024. DOI: 10.1109/TCSS.2025.3588779.
[24] Yan, J.; et al. Multimodal Emotion Recognition Based on Facial Expressions, Speech, and Body Gestures. Electronics 2024, 13 (18), 3756. DOI: 10.3390/ELECTRONICS13183756.
[25] Singh, J.; Bhangu, K. S.; Alkhanifer, A.; AlZubi, A. A.; Ali, F. Quantum Neural Networks for Multimodal Sentiment, Emotion, and Sarcasm Analysis. Alexandria Eng. J. 2025, 124, 170–187. DOI: 10.1016/J.AEJ.2025.03.023.
[26] Liu, Y.; Li, Q.; Wang, B.; Zhang, Y.; Song, D. A Survey of Quantum-Cognitively Inspired Sentiment Analysis Models. ACM Comput. Surv. 2023, 56 (1). DOI: 10.1145/3604550.
[27] Li, Z.; Zhou, Y.; Liu, Y.; Zhu, F.; Yang, C.; Hu, S. QAP: A Quantum-Inspired Adaptive-Priority-Learning Model for Multimodal Emotion Recognition. In Proceedings of the Annual Meeting of the Association for Computational Linguistics; 2023, 12191–12204. DOI: 10.18653/V1/2023.FINDINGS-ACL.772.
[28] Yi, M. H.; Kwak, K. C.; Shin, J. H. HyFusER: Hybrid Multimodal Transformer for Emotion Recognition Using Dual Cross Modal Attention. Appl. Sci. 2025, 15 (3), 1053. DOI: 10.3390/APP15031053.
[29] Krouska, A.; et al. Meaningful Multimodal Emotion Recognition Based on Capsule Graph Transformer Architecture. Information 2025, 16 (1), 40. DOI: 10.3390/INFO16010040.
[30] Sun, S.; Cao, R.; Rutishauser, U.; Yu, R.; Wang, S. A Uniform Human Multimodal Dataset for Emotion Perception and Judgment. Sci. Data 2023, 10, 773. DOI: 10.1038/S41597-023-02693-Z.