Library Progress International *Print version ISSN 0970 1052*
Vol.44 No. 3, Jul-Dec 2024: P.12675-12684 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Multi-Model Emotion Recognition from voice, face, and text sources using optimized Equivariant Quantum Dense Nested Neural Networks

K. Murugesan[1], Dr. S. Sivakumar[2], Dr. J Venkatesh[3], S. Balaguru[4] and N. Revathi[5]

[1]Associate Professor, Department of Electrical and Electronics Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
Orcid id: (0000-0003-1977-4914)
[2]Assistant Professor, Department of Computer Applications, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
Orcid id: (0009-0002-4432-6646)
[3]Professor, Department of Computer Science and Engineering, Chennai Institute of Technology, Kundrathur, Chennai, India
Orcid id: (0000-0002-4259-130X)
[4]VIT Bhopal University, Kothrikalan, Sehore-466114
Orcid id: (0000-0002-1160-4446)
[5]Assistant Professor, Department of ECE, Panimalar Engineering College, Chennai, India
[1]murugesank@ssn.edu.in, [2]sivaskumarmca@gmail.com, [3]venkateshj@citchennai.net, [4]balaguruiitm@gmail.com and [5]revathinandhan@gmail.com

How to cite this article: K. Murugesan, S. Sivakumar, J Venkatesh3, S. Balaguru, N. Revathi (2024). Multi-Model Emotion Recognition from voice, face, and text sources using optimized Equivariant Quantum Dense Nested Neural Networks *Library Progress International,* 44(3), 12675-12684.

ABSTRACT
The term Emotion Recognition(ER) is used for the Recognition and comprehension of human emotions have become essential aspects of technological progress in today's fast-paced, linked world where human–computer interaction is an essential part of everyday life. It is necessary to identify ER from audio,video and text data, because of that many techniques are implemented. However,the existing methods have lack of accuracy, precision and high error rate. To overcome the aforementioned problem, Equivariant Quantum Dense Nested Neural Networks(EQDNNN) with Fire Hawk Optimizer (EQDNN-FHO) is proposed for accurately identifying ER from audio, video and text data. In this input image is taken from two datasets such as IEMOCAP dataset and RAVDESS dataset. To improve the quality of the speech signals, text, and video signals, undesired artifacts are eliminated using methods including tokenization for text, Spectral Noise Gate (SNG) for audio and Color Wiener Filtering (CWF) for video. After that, the features such as audio, video and text are extracted using Short-Time Fourier Transform, Video Feature Extraction and Bag-of-Words is used for subsequent extraction. After that classification are done using Equivariant Quantum Dense Nested Neural Network (EQDNN)and optimization are done using Fire Hawk Optimizer (FHO) for detecting the different types of actions done by humans. The efficiency of the proposed EQDNN -FHO is analyzed using a dataset and attains 99.7% accuracy, 99.27% recall and attains better results compared with the existing methods. This method performed similarly in identifying positive and negative emotions when compared to human annotation. Moreover, it has demonstrated efficacy in precisely identifying the sentiment of ambiguous emotion parts that were deemed unsuitable in prior research. Advances in audio emotion recognition may have consequences for voice-activated virtual assistants, medical fields, and other industrial uses of automated communication.

Keywords: Multi-Model Emotion Recognition, Spectral Noise Gate, Color Wiener filtering, tokenization, Equivariant Quantum Dense Nested Neural Network, Fire Hawk Optimizer

1. INTRODUCTION

Human-robot interaction, marketing, and technology are just a few of the domains where emotion recognition is essential. Joy, love, surprise, sadness, anger, and fear are among the six categories into which Ekman's model divides emotions. Empathy detection from text, speech, and facial expressions is known as Automatic Emotion Recognition (AER). To overcome the shortcomings of single-source methods and produce more accurate identification, multimodal ER processing makes advantage of many information sources, including facial

motions. For the past fifty years, there has been a noticeable increase in interest in this field[1-3]. Recognition of emotions is important in several domains, including assisted driving, psychiatric counseling, and remote learning. Accurate emotion identification still faces challenges, though. Due to the common emotional drive and cross-coupling of human social activities, multimodal fusion provides a solution for accurate emotion recognition. Representation learning and fusion mechanisms are key components of the primary problems for multimodal learning [4-6]. Emotions are intricate mental and physical states that represent a person's mental state. They can be characterized as complex experiences of consciousness, physical sensation, and action expressing personal significance, or as intense emotions such as happiness, grief, wrath, or surprise. It is possible to think of emotions as processes or states that interact with other mental states to produce particular behaviors. They are defined as deliberate mental responses, typically focused on a particular item, that are followed by alterations in the body's physiology and behavior. Emotions and sentiments can be efficiently conveyed through written, auditory, and visual channels in human communication, which is intrinsically multimodal. Accurate emotion identification requires integrating several senses.[7-9]. AI applications and human-computer interface depend on the accurate recognition of human emotions in dialogue. Multiple modalities must be supported by AI technologies, according to research, in order for them to function well in practical settings. Although good multi-modal fusion approaches have been developed, uni-modal emotion recognition is still necessary for certain applications. Feelings are individualized, hence it might be difficult to read other people's feelings. With the introduction of audio modality in talks, this study aims to improve the accuracy of Speech Emotion Recognition (SER) by encoding audio cues that narrate emotions. Since deep learning models are used in current methods, interpretability is restricted[10-12].

Novelty and Contribution

The Novelty and contribution of this paper is given below:

- In this manuscript, a Equivariant Quantum Dense Nested Neural Networks(EQDNNN) with Fire Hawk Optimizer (EQDNN-FHO) based Deep Feature Vectors Concatenation for (DDGANN-DKAN-GAO) is proposed
- To improve the quality of the speech signals, text, and video signals, undesired artifacts are eliminated using methods including tokenization for text, Spectral Noise Gate (SNG) for audio and Color Wiener filtering (CWF) for video.
- After the pre-processing the features such as audio,video and text are extracted using enhanced Gaussian short-time Fourier transform (STFT) spectrograms and continuous wavelet transform,video feature extraction and Bag-of-Words.
- Feature fusion and Classification are done using Equivariant Quantum Dense Nested Neural Networks Based Feature Fusion And Emotion Recognition (EQDNNN) it is used for classify the positive negative and neutral emotions.And Optimization are done using Fire Hawk Optimizer is used to optimize the weight parameters of EQDNNN-FHO for improving accuracy .

2. LITERATURE SURVEY:

In 2023 Dal Rì et.al[13] has introduced an Convolutional Attention Block and CNN-based model for Speech emotion recognition(SER). This method is applied to four widely-used English datasets for SER applications, although deep learning has led to major progress in Speech Emotion Recognition (SER), which challenges remain due to the lack of established best practices and high-quality datasets.

In 2023 Guizzo et.al[14] has introduced a ResNet-50, for speech emotion identification .In this method the tasks are used for the purpose of extracting quaternion embeddings from realvalued monoaural spectrograms. It is composed of a quaternion-valued decoder, a real-valued encoder, and a real-valued emotion classifier which is Tested on four widely-used datasets, ResNet-50's performance is compared against three well-known real-valued architectures and their quaternion-valued counterparts. The outcomes demonstrate a steady increase in test accuracy with a decrease in model resources.

In 2023 Voloshina et.al[15] has introduced an Bidirectional Encoder Representation Transformer (BERT) for improving emotion recognition. This method combines multimodal interaction and disguised multimodal attention with textual, audio, and video data. This method is fine-tuning for textual, audio, and video input, based on the language representation paradigm.

In 2024 Gómez-Zaragozá, Lucía et.al[16] has introduced a Unispeech-SAT-Large(USSATL) model for emotion recognition prediction. The Emotional Voice Messages (EMOVOME) database in this method yielded the best results, which contains 999 audio messages from actual interactions with 100 Spanish people. The research assessed speaker-independent SER models and contrasted its findings with databases of reference. The pre-trained UniSpeech-SAT-Large model

In 2024 Wang et.al[17] has introduced an BLSP-Emo method, for emotions and speech recogniition. It is an revolutionary method for creating an end-to-end speech-language model and produces sympathetic output, which have the capacity to produce expressive speech with a wide range of emotions and low latency, as demonstrated by the GPT-4o release. For semantic alignment and emotion alignment, BLSP-Emo leverages speech and emotion recognition datasets.

In 2022 Abdelhamidet.al[18] has introduced an Deep Neural Network (DNN) for speech emotion recognition. This technique reduces noise fractions and optimizes hyperparameters to improve the,where the stochastic fractal search (SFS)-guided whale optimization algorithm (WOA) is used in this method to maximize the learning rate and label smoothing regularization factor. To provide the best possible global solutions, the suggested method is evaluated using the IEMOCAP, Emo-DB, RAVDESS, and SAVEE speech emotion datasets.

Problem Statement:

Current audio, video, and text-based emotion recognition (ER) techniques have poor precision, massive error rates, and low accuracy. IEMOCAP and RAVDESS datasets are used for evaluation while recommending Equivariant Quantum Dense Nested Neural Networks with Fire Hawk Optimizer (EQDNN-FHO) for enhanced ER performance in order to address these problems.

3 PROPOSED METHODOLOGY

Figure 1 illustrates the EQDNNN-FHO operation. The three steps of the suggested method are: (1) pre-processing (2) Feature extraction (3) Feature fusion and classification (4) Optimization

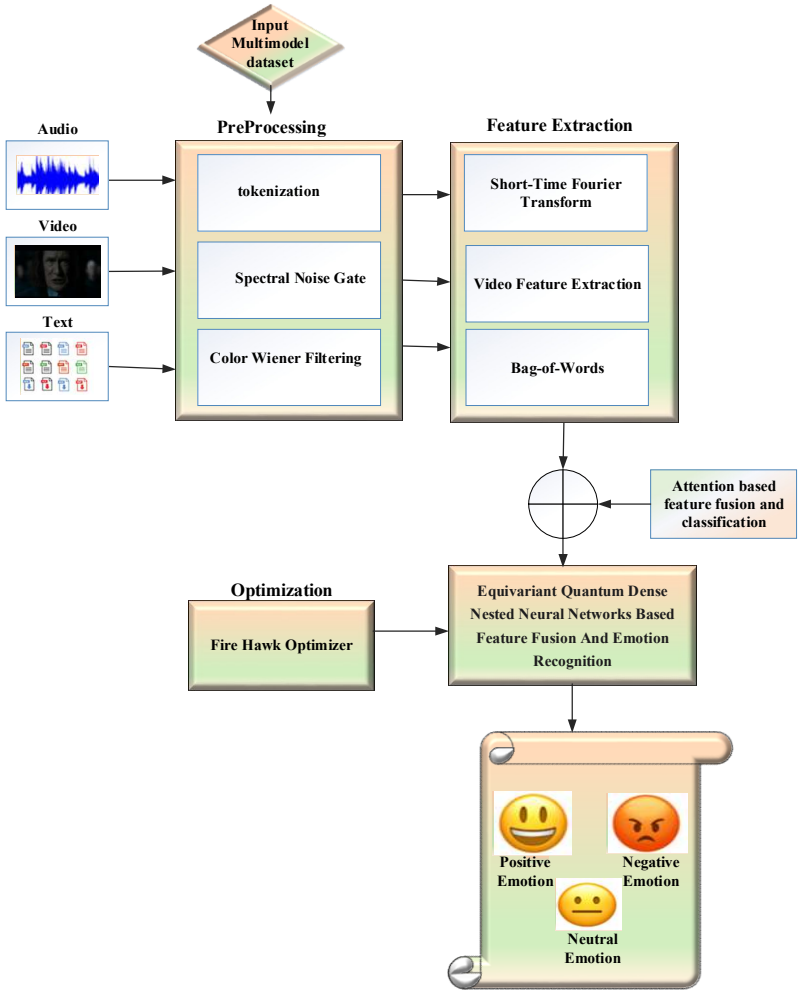


Figure 1: Workflow diagram of proposed EQDNNN-FHO method

3.1 Image Acquisition

In this input data is taken from audio signals, video signals and Text using two datasets such as IEMOCAP dataset and RAVDESS dataset. Hence, these data are full of noises, to remove these noises pre-processing

techniques are applied in three signals such as audio signal, video signal and text are used, they are mentioned in given below:

3.2 Pre-processing

The Comprehensive pre-processing is the first step in preparing the audio, video data and text. This step is essential to reducing the noise, incompleteness, and consistency problems that often arise with raw data. To improve the quality of the speech signals, text, and visual frames, undesired artifacts are eliminated using methods including tokenization for text, Spectral Noise Gate (SNG) for audio and Color Wiener filtering (CWF) for video

- **Silence removal in audio:**

Controlling audio noise is essential for guaranteeing, so that Spectral Noise Gate method is used[19-20]**.** Silence can be heard in large portions of the audio transmission. The lack of information in the quiet signal makes it meaningless. Zero crossing rate (ZCR) and short time energy (STE) are two techniques that can be used to remove these components. The frequency at which the voice signal's amplitude falls to zero during a specified time period is known as the "zerocrossing rate." The following equation[1] can be used to calculate the energy of a signal during a given time interval.

$$B_h = \sum_{n=-\infty}^{\infty} \left[y(n) s(m-n) \right]^2 \qquad (1)$$

Where, B_h denotes the result of the convolution at position h $y(n)$and $s(m-n)$denotes the sequences and m, n denoted as indices. The process of removing silence involves calculating the highest value for every frame and comparing it to a predetermined threshold. A frame is inserted as a speech frame if its maximum value above the threshold; if not, it is removed as a quiet frame.

Color Wiener filtering (CWF) for video:

The noises in the video signals are removed using Color Wiener filtering (CWF)[21]. The extraction of local shape and texture information from video material is facilitated by the use of histograms of directed gradient models. A deeper comprehension of the visual data is facilitated by the use of color Wiener filtering to extract contextual information from the frames. An inverse filtering method called color Wiener filtering can be used to sharpen blurry photos, however it might not be able to maintain structural details in scenes from nature or sporting events. The following equation(2) relates recovered color space and additive noise in the RGB original color space,

$$R_j = S_j + \lambda_j \qquad (2)$$

Where, R_j denotes the recovered or observed image that corresponds to the input image's pixel vector S_j and λ_j represents the addictive noise.

- **Tokenization**

The text data are preprocessed using the method called Tokenization[22]. The process of converting text into a list of tokens is known as tokenization. These tokens are fed into machine learning models after being encoded with integers. Tokenization can be achieved by segmenting the text into terms with inherent meaning, where white spaces can be readily used. But there are drawbacks to utilizing words as tokens, such as vocabulary size issues that may arise depending on the size of the training dataset. Typically, this difficulty is solved by decreasing the vocabulary size, but this leaves many unknowns, particularly in languages with rich morphology. A well-designed tokenization strategy separates tokens with white space and characters at a balanced granularity. Splitting words into units known as subwords is made possible by subword tokenizations. The capacity to generate subwords not only restricts the vocabulary size but also enables the construction of universal tokenization across many languages. Then the pre-processed data is given to the feature extraction.[23]

Feature extraction:

After preprocessing the feature extraction is carried out.The features such as audio,video and text are extracted using enhanced Gaussian short-time Fourier transform (STFT) spectrograms and continuous wavelet transform,video feature extraction and Bag-of-Words.

- **Short-Time Fourier Transform(STFT) for audio:**

The features of audio is extracted using Short-Time Fourier Transform(STFT)[24]. By applying the Fourier Transform to brief, overlapping signal segments, the Short-Time Fourier Transform (STFT) method provides

time-localized frequency information for assessing the frequency content of non-stationary signals. A one-dimensional matrix $h[j,\sum\ ]$ is produced by the STFT operation, which accepts a one-dimensional signal $h(T)$ as input. A tapering function gL of length L multiplied by an index $T_j + M - 1$ of the signal s's Discrete Fourier Transform (DFT) of a slice of length L is represented by each column $h[j:]$ of the STFT. In mathematics, STFT can be expressed in equation(3) as follows:

$$h[j,e] = \sum_{P=0}^{M-1} om[p]h[T_j + p]f^{-2i\pi p/M} \quad (3)$$

Where, $h[j,e]$ denotes the output of the transform for j th window, M denoted as the length of the window, p denotes as the index, $om[p]$ denotes as the signals window function, $h[T_j + p]$ represents as the signal value at the j th window and p th sample, T_j represents the start time.

- **Video Feature Extraction:**

The features of video are extracted using the Video Feature Extraction[25]**.** A video is an image sequence that has temporal properties and is more informative than a series of still photos. The video feature extaction is chosen in order to enhance face detection performance in the video stream. The following are the steps in the method:

1. The video image is extracted within a predetermined time frame. Since IEMOCAP requires a minimum sample segmentation unit of 100 ms, this unit is chosen as the fundamental unit of sample segmentation.
2. From the video image, the coarse-grained area containing the head is extracted.
3. The face is represented as a Haar-like feature, and the video features are swiftly calculated using an integral graph.
4. Rectangular features (weak classifiers) that best capture the face are chosen using the video feature extraction process, and then the weak classifiers are combined into a strong classifier by weighted voting.
5. To enable the mapping of the video segment to the speech segment, the video is recut while keeping the facial portion of the picture in place based on its position in the image.The final size of the facial image is 70 by 70.

- **Bag-of-Words for text:**

The features of text are extracted using the method called Bag-of-Words(BoW)[26]. This approach is popular and highly effective in identifying and categorizing the attributes by making bags for every kind of instance. Information retrieval (IR) and natural language processing (NLP) both use the BoW model, a condensed representation. This model treats texts, such as sentences or documents, as a bag of words; it ignores word order and syntax in favor of merely looking at word duplication. The BoW model is mostly utilized in document classification techniques, where each word's frequency or occurrence generates the features needed to train a classifier. Because of its versatility across several domains, it can be applied to a wide range of situations. Among those goals are:

1. Text classification:

Organizing texts into classes or categories and determining the weights assigned to each word based on how frequently it appears in the text.

2. Image Recognition and Classification:

Categorizing visual scenes (pictures and videos) by creating methods that can capable to Classify visual scenes, Detect and localize objects, Estimate semantic and geometrical attributes, Classify human activities and events.

The features such as audio,video and texts are extracted and given to the fusion for classifying the emotions.

EQUIVARIANT QUANTUM DENSE NESTED NEURAL NETWORKS BASED FEATURE FUSION AND EMOTION RECOGNITION

After the extraction of features from video, audio and text data, it combines Equivariant Quantum Neural Networks with Dense Nested Attention Network for the purpose of fusion and then it classify the emotions respectively.

- **Equivariant Quantum Neural Networks**

Equivariant Quantum Neural Networks (EQNNs)[27] are quantum neural networks engineered to maximize

efficiency and performance, By utilizing data symmetries in quantum calculations. Symmetry transformations operating on the input feature embedding space in EQNN models are implemented as finite-dimensional unitary transformations $V_f, f \in F$ Assume about the most basic scenario in which there is a single trainable operator $V(\eta, y)$ acts on a state of $V(\eta, y)/\phi>$.For the purpose of symmetry transformation V_f the condition is given in equation(4)

$$V(\eta, y)V_f /\phi> = V_f V(\eta, y)/\phi> \quad (4)$$

where $V(\eta, y)$ represents the unitary operator, V_f denotes the quantum circuit gate operator, $/\phi>$ represents the quantum state key notation, $V_f /\phi>$ denotes the state $/\phi>$ after the unitary operator application V_f.

- **Dense Nested Attention Network**

This innovative method aims to categorize emotions holistically in addition to promoting a thorough understanding of emotions by utilizing the synergistic potential of various data sources[28]. A neural network architecture known as a Dense Nested Attention Network (DNANet) combines nested attention mechanisms and dense connections to improve feature extraction and learning, hence enhancing performance in applications such as natural language processing and picture recognition. This approach works particularly well for expressing the complex range of human emotions. As a result, the issue is changed to create a multiple-attribute categorization model, which has the following equation(5)

$$x = F_\Theta(y_h), h \in [M] \quad (5)$$

where, x represents the output, F_Θ denotes the enhanced DNA-net, y_h denotes the input data at the index h, $h \in [M]$ where h is the element of the set $[M]$. The Channel and Spatial Attention Module then processes each node in DNIM, which can improve a tiny target's features. Convolutional neural networks use the Channel Attention Module to handle channel-based attention is given in equation(6)

$$N_d(M) = \beta(MLP(R_{avg}(M)) + MLP(R_{\max}(M)) \quad (6)$$

Where, β denotes the sigmoid function, $(R_{avg}(M)$ represents the average value of some property R, $(R_{\max}(M)$ denotes the average pooling and max-pooling operations, MLP denotes the A kind of neural network called a multi-layer perceptron is utilized for tasks involving regression or classification.

- **Loss Function**

The binary crossentropy loss based on focal loss and dice loss based on the dice coefficient make up the two parts of the loss function in the Enhanced DNA-net. As the total loss for the binary classification task, we sum them all together given in equation(7)

$$H_{total} = \alpha H_{dh} + (1-\alpha)H_{gh} \quad (7)$$

where, α denotes the weighting factor, H_{gh} denotes the loss for categorizing the pixels in which the vessels H_{dh} represents the decrease of similarity between the actual segmentation results and the anticipated segmentation, H_{total} denotes the weighted sum of two distinct loss functions, often known as the objective function.To preserve the sensitivity of H_{dh} to small targets while utilizing H_{dh} capacity to handle sample imbalance by performing a weighted average of dice loss and focused loss by varying the value of α. To be more precise, the classification model separates emotions into three main categories.

a. Positive emotion

Feelings like happiness, enthusiasm, and an overall sensation of optimism are among them. These feelings are characterized by being positive and affirming.

b. Negative emotion

Emotions such as sadness, fear, and disgust fall under this group since they are usually connected to negative experiences or impressions.

c. Neutral Emotion

This category, which doesn't clearly fit into either of the other two categories, represents a state that is neither overtly positive nor negative.

The EQDNNs can take advantage of both the advantages of dense nested neural networks and the advantages of quantum computing, all while preserving symmetry and invariance qualities, they can be a potent method for emotion recognition. Hence the combination of Equivariant Quantum Neural Networks with Dense Nested Attention Network is used for the purpose of fusion and then it classify the emotions such as positive, negative and neutral respectively.

3.4.1 Optimization using Fire Hawk Optimizer (FHO)

Fire has long been a component of the cultural and ethnic traditions of the native Australians, who use it to maintain and control the equilibrium of the surrounding ecosystem and terrain [29].Fires that start spontaneously or are purposefully ignited can spread due to people and other reasons triggered by lightning, increasing the biodiversity and local ecosystem's susceptibility. Moreover, it was recently discovered that brown falcons, whistling kites, and black kites are capable of starting fires that spread over the area. According to recent studies, brown falcons, whistling kites, and black kites can start fires. The aforementioned birds, also called Fire Hawks, deliberately spread fire by carrying blazing sticks in their beaks and talons, a behavior that is described as a natural disaster. The birds pick up flaming sticks. Equation(8) which changes according on the task, typically indicates the goal function that needs to be minimized,

$$Fitness\,\text{Function} = \text{Maximize}(\alpha)$$

Where, α is used for improving accuracy. A metaheuristic algorithm called the Fire Hawk Optimizer was developed in response to fire hawks' hunting techniques. By mimicking predator-prey interactions, it strikes a balance between exploration and exploitation to effectively solve challenging optimization issues.

4 RESULT AND DISCUSSIONS

This section describes the introduced scheme's findings and debate. Python is used to carry out the Result and discussion of the method. Here is some of the Implementation parameter is mentioned in table 1:

Table 1: Implementation Parameters

Parameters	Description
Proposed Neural Network	EQDNNN-FHO
OS	Windows 10
Optimization	FHO
Dataset	IEMOCAP dataset ,Ravdess Dataset
Software	Python 3.7

4.1 Dataset Description

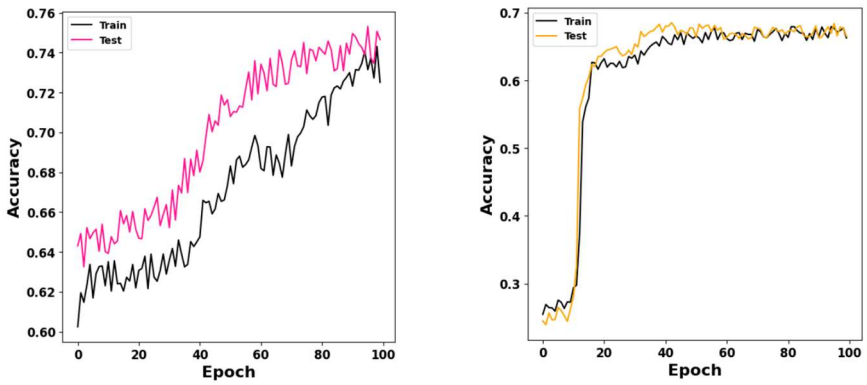
To classify the ER, an ER dataset is taken; namely IEMOCAP dataset and RAVDESS dataset and its descriptions are given below:

4.1.1 IEMOCAP dataset

A 12-hour dataset called Interactive Emotional Dyadic Motion Capture (IEMOCAP) includes audiovisual recordings from five sessions in which actors performed both scripted and spontaneous speech[8]. The distribution of the dataset's samples, which come from nine different emotional categories, is uneven. In order to solve this, the dataset was preprocessed, yielding a database of 7664 samples from four different emotional classes by slicing lengthy samples into smaller chunks.

4.1.2 Ravdess dataset

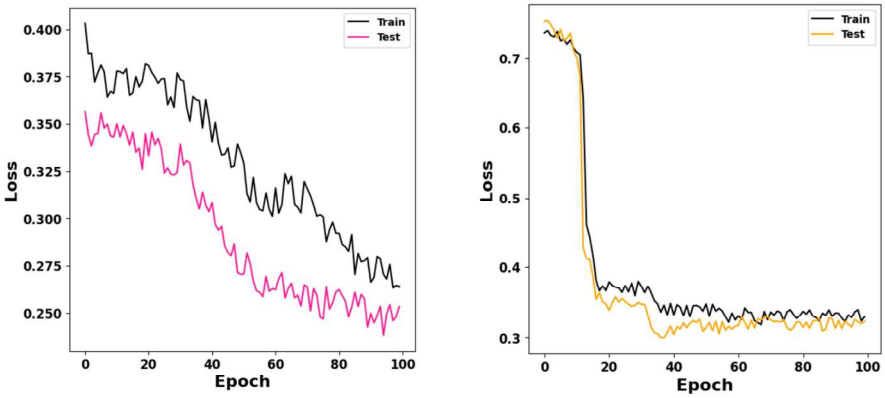
The 7356 audiovideo recordings of 24 professional actors, comprising 1440 utterances and eight emotion classes (neutral, calm, happiness, sorrow, anger, fear, surprise, and disgust),where happiness and surprise are comes under positive emotion ,calm, sorrow, fear and disgust are comes under negative emotion,the final class is known as Neutral emotion respectively in the RAVDESS dataset, a consolidated standard for ER applications[8]. It provides high-quality audio and video recordings of actors portraying different emotions, which can be used to train and test algorithms for emotion recognition in both speech and song.



a b

Figure 2: Training Accuracy for (a) IEMOCAP dataset (b) RAVDESS dataset

Figure 2 shows the Training Accuracy for (a) IEMOCAP dataset (b) RAVDESS dataset with 99.7% accuracy. By evaluating the model's performance on the training set, training accuracy offers valuable information about how well the model is assimilating the material it has encountered.



a b

Figure 3: Loss function for (a) IEMOCAP dataset (b) RAVDESS dataset

Figure 3 shows the Loss function for (a) IEMOCAP dataset (b) RAVDESS dataset. One important measure of the model's learning efficiency during training is the loss function. A lower loss function value indicates greater model performance. Table 2 shows the Comparison for performance analysis of proposed approach.

Table 2: Comparison for performance analysis of proposed approach

Methods	Dataset	Acc (%)	Pre (%)	Spe (%)	Recall (%)	PT (%)	Error (%)
CNN[8]	**IEMOCAP dataset and RAVDESS dataset**	98.3	96.7	97.5	97.89	0.8	0.34
ResNet-50[9]	**IEMOCAP dataset and RAVDESS dataset**	88.6	90.2	87.9	83.5	0.7	0.26
BERT[10]	**IEMOCAP dataset and RAVDESS dataset**	91.8	90.23	89.97	89.12	0.9	0.33
USSATL[11]	**IEMOCAP dataset and RAVDESS dataset**	87.5	88.6	93.5	86.68	0.32	0.25
BLSP-Emo[12]	**IEMOCAP dataset and RAVDESS dataset**	90.6	87.57	89.65	84.56	0.10	0.41
DNN[13]	**IEMOCAP dataset and RAVDESS dataset**	91.4	84.5	87.5	88.4	0.6	0.26
EQDNNN-FHO	**IEMOCAP dataset and RAVDESS dataset**	**99.7**	**98.97**	**99.4**	**99.27**	**0.1**	**0.10**

Here, Acc signifies Accuracy, Pre signifies precision, Spe signifies the Specificity,PT signifies processing time .

The above Table 2 illustrations the qualified analysis of the proposed method against existing methods CNN[8], ResNet-50[9], BERT[10] USSATL[11], BLSP-Emo[12],DNN[13]methods through the IEMOCAP dataset and RAVDESS dataset. The proposed method attains the highest accuracy (99.7%), precision (98.97%),Spe(99.4%),Recall(99.27%). It likewise validates the lowest processing time (0.1) error rate (0.10), demonstrating its efficiency and reliability.

5. CONCLUSION

In this manuscript, the input data is taken from a dataset such as IEMOCAP dataset and RAVDESS dataset. To improve the quality of the speech signals, text, and video signals, undesired artifacts are eliminated using methods including tokenization for text, Spectral Noise Gate (SNG) for audio and Color Wiener Filtering (CWF) for video. After that, the features such as audio, video and text are extracted using Short-Time Fourier Transform, Video Feature Extraction and Bag-of-Words is used for subsequent extraction. Then the classification and optimization are done using EQDNNN-FHOfor identifying the ER. The introduced system is executed in python. The efficiency of the proposed EQDNNN-FHOis analyses using IEMOCAP dataset and RAVDESS dataset attain 99.7% accuracy and 0.10% error rate, compared with the existing methods. Moreover, it has demonstrated efficacy in precisely identifying the sentiment of ambiguous emotion parts that were deemed unsuitable in prior research. Advances in audio emotion recognition may have consequences for voice-activated virtual assistants, medical fields, and other industrial uses of automated communication.

REFERENCES

1. Fu, Rui, et al. "MM Dialogue GAT-A Fusion Graph Attention Network for Emotion Recognition using Multi-model System." *IEEE Access* (2024).
2. M. Preetha, Archana A B, K. Ragavan, T. Kalaichelvi, M. Venkatesan “A Preliminary Analysis by using FCGA for Developing Low Power Neural Network Controller Autonomous Mobile Robot Navigation”, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), ISSN:2147-6799. Vol:12, issue 9s, Page No:39-42, 20245184.
3. Wu, Junyan, et al. "Audio Multi-view Spoofing Detection Framework Based on Audio-Text-Emotion Correlations." *IEEE Transactions on Information Forensics and Security* (2024).
4. Balaji Singaram, Lakshmi. B, Dr.M.Preetha, V.K. Ramya Bharathi, Dr.S.Muthumari lakshmi, Rakesh Kumar Giri “A Smart IoT-Based Fire Detection and Machine Learning Based Control System for Advancing Fire Safety”,Nanotechnology Perceptions, ISSN 1660-6795 2024, Vol: 20, 5s, 229-244.
5. K Siva Kumar, T.Mayavan, S. Sambath, A. Kadirvel, T.S.Frank Gladson, S. Senthilkumar ”Artificial Neural Networks to the Analysis of AISI 304 steel sheets through limiting Drawing Ratio test” Journal of Electrical Systems 2024, Vol: 20, 4s, 2282-2291.
6. M .Preetha, K. Monica, H. Nivetha, S. Priyanka “A Survey on Semi-Supervised Image to Video Adaptation for Video Action Recognition”, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol No: 6, Issue No: 2, Page No: 2513-2517, February 2018. ISSN: 1748-0345.
7. Xu, Yurui, et al. "A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context." Complex & Intelligent Systems 9.1 (2023): 951-963.
8. M. Preetha, Raja Rao Budaraju, Jackulin. C, P. S. G. Aruna Sri, T. Padmapriya “Deep Learning-Driven Real-Time Multimodal Healthcare Data Synthesis”, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), ISSN:2147-6799, Vol.12, Issue 5, page No:360-369, 2024.
9. Priyadarshinee, Prachee, et al. "Alzheimer’s dementia speech (audio vs. text): Multi-modal machine learning at high vs. low resolution." *Applied Sciences* 13.7 (2023): 4244.
10. Balaji Singaram, M.S.Vinmathi, Dr.H.B.Michael Rajan, Jeyamohan H, T. Manikandan, “Data-Driven Estimation of Lithium-Ion Battery State-of-Health Prediction Approach Using Machine Learning Algorithm for Enhanced Battery Management Systems”, Nanotechnology Perceptions, ISSN 1660-6795 2024, Vol: 20, 7s, 93-103.
11. K Sivakumar, M. Sughasiny, K.K.Thyagarajan, A. Karthikeyan & K. Sangeetha “A Comparative Analysis of GOA (Grasshopper Optimization Algorithm) Adversarial Deep Belief Neural Network for Renal Cell Carcinoma: Kidney Cancer Detection & Classification,” International Journal of Intelligent Systems and Applications In Engineering, ISSN: 2147-6799, 2024, 12(9s), 43–48.

12. Khan, Umair Ali, et al. "Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects." *Multimedia Systems* 30.3 (2024): 115.

13. Dal Rì, Francesco Ardan, Fabio Cifariello Ciardi, and Nicola Conci. "Speech Emotion Recognition and Deep Learning: an Extensive Validation using Convolutional Neural Networks." *IEEE Access* (2023).]

14. Guizzo, Eric, et al. "Learning speech emotion representations in the quaternion domain." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023): 1200-1212

15. Voloshina, Tatiana, and Olesia Makhnytkina. "Multimodal Emotion Recognition and Sentiment Analysis Using Masked Attention and Multimodal Interaction." *2023 33rd Conference of Open Innovations Association (FRUCT)*. IEEE, 2023.

16. Gómez-Zaragozá, Lucía, et al. "Speech emotion recognition from voice messages recorded in the wild." arXiv preprint arXiv:2403.02167 (2024).

17. Wang, Chen, et al. "BLSP-Emo: Towards Empathetic Large Speech-Language Models." *arXiv preprint arXiv:2406.03872* (2024).

18. Abdelhamid, Abdelaziz A., et al. "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm." *Ieee Access* 10 (2022): 49265-49284.

19. Nema, Bashar M., and Ahmed A. Abdul-Kareem. "Preprocessing signal for speech emotion recognition." Al-Mustansiriyah Journal of Science 28.3 (2018): 157-165.

20. I. A. Karim Shaikh, P. Jagdish Patil, R. Sethumadhavan, M. Preetha and H. Patil, "Predictive Analysis of Employee Turnover in IT Using a Hybrid CRF-BiLSTM and CNN Model," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 914-919, doi: 10.1109/ICSCNA58489.2023.10370093.

21. Kumar, MP Pavan, et al. "Structure-preserving NPR framework for image abstraction and stylization." The Journal of Supercomputing 77.8 (2021): 8445-8513….video

22. Alyafeai, Zaid, et al. "Evaluating various tokenizers for Arabic text classification." Neural Processing Letters 55.3 (2023): 2911-2933

23. Dr.M.Preetha, Balaji Singaram, Dr.I. Manju, B.Hemalatha, P. Bhuvaneswari "Machine Learning in Breast Cancer Treatment for Enhanced Outcomes with Regional Inductive Moderate Hyperthermia and Neoadjuvant Chemotherapy" Nanotechnology Perceptions, ISSN 1660-6795 2024, Vol: 20, 5s, 245-259.

24. Balazs, Peter, et al. "Comparisons between Fourier and STFT multipliers: The smoothing effect of the short-time Fourier transform." Journal of Mathematical Analysis and Applications 529.1 (2024): 127579..

25. Jia, Ning, Chunjun Zheng, and Wei Sun. "A multimodal emotion recognition model integrating speech, video and MoCAP." Multimedia Tools and Applications 81.22 (2022): 32265-32286.

26. Opitz, Juri. "Gzip versus bag-of-words for text classification with KNN." arXiv preprint arXiv:2307.15002 (2023).

27. Dong, Zhongtian, et al. "$\mathbb{Z}2\times \mathbb{Z}2$ Equivariant Quantum Neural Networks: Benchmarking against Classical Neural Networks." Axioms 13.3 (2024): 188.

28. Zuo, Gao, et al. "Night-Time Vessel Detection Based on Enhanced Dense Nested Attention Network." Remote Sensing 16.6 (2024): 1038.

29. Shishehgarkhaneh, Milad Baghalzadeh, et al. "BIM-based resource tradeoff in project scheduling using fire hawk optimizer (FHO)." *Buildings* 12.9 (2022): 1472.