https://doi.org/10.48047/AFJBS.6.14.2024.7540-7569

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Multimodal Emotion Recognition Using Novel Feature fusion mechanism with SVM

Sanjeeva Rao Sanku[1],B. Sandhya[2]

[1]Research Scholar, CSE department,University College of Engineering, Osmania University, India
E-mail: ssanjeevarao@gmail.com
[2]Professor, CSE department, MVSR Engineering College, India, E-mail: sandhya_cse@mvsrec.edu.in

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.7540-7569

Abstract

Emotion recognition based multimodal data (e.g., video, audio, text, etc.) is a highly demanding and significant research field with numerous applications. This research rigorously explores model-level fusion to discover the best multifunctional model combining audio and visual modalities for emotion identification. Specifically, it proposes novel feature extractor networks for both video and audio data. This research presents a comprehensive approach to multimodal emotion recognition, utilizing state-of-the-art feature extraction methods tailored to each modality. For text data, we implement the Assimilated N-gram Approach (ANA) to effectively capture contextual information. Audio features are extracted using Mel-Frequency Cepstral Coefficients (MFCC), ideal for capturing spectral characteristics in speech. Visual features are derived using SqueezeNet, a deep learning architecture optimized for efficient and informative visual data representation. To integrate the extracted features from text, audio, and visual modalities, we propose a multimodal data fusion strategy that combines information across modalities, thereby enhancing the overall representation of emotional cues. In the classification stage, we employ Support Vector Machine (SVM), a robust and effective classification algorithm known for its ability to handle high-dimensional data and perform well in diverse scenarios. Using the Multimodal Emotion Lines Dataset (MELD), our approach achieved an accuracy of 98.1%, precision of 98.85%, recall of 98.75%, and F-measure of 98.3. These results highlight the effectiveness of our multimodal framework in emotion recognition tasks.

Keywords: Assimilated N-gram Approach, Mel-Frequency Cepstral Coefficients, Squeezenet, Support Vector Machine

1. Introduction

Any conscious experience that is marked by heightened mental activity and a certain level of pleasure or pain is considered emotional [1]. Emotional intelligence enables the human-machine interaction more harmoniously [2]. A significant field of study at the nexus of artificial intelligence and interpersonal communication analysis is recognizing emotions. It is really useful

for robots [3]. Emotions are very important in human decision handling, interaction and cognitive process [4]. Speech emotion recognition is one of key components for human-computer interaction systems [5]. Emotions are present in almost every decision and moment of our lives. Thus, recognizing emotions awakens interest, since knowing what others feel lets us interact with them more effectively [6].

Modern detectors that can record both audio and visual signals are opening the door for a host of cutting-edge innovations that will enable discreet, contact-free monitoring and diagnostics [7]. Deep learning methods have recently shown effective in solving issues in a number of domains, including text-to-speech creation, picture categorization, translation by machine, and speech recognition [8]. Humans may exhibit emotion using a variety of modalities, including speech, body language, and facial expressions. As a consequence, using many modalities can help recognize emotion more correctly than unimodal methods [9]. The need to develop automated emotion identification systems is rising in tandem with the development of technology and our expanding knowledge of emotions. Speech has been used in several studies on emotion identification [10].

The restricted amount of emotion information that may be found in a single mode represents a few of the difficulties in recognizing emotions. It is challenging to fulfill the requirements of the existing emotional identification system to obtain the correct emotional state simply from a single modalities because to the proliferation in audio information on social media [11]. The process of multimodal emotion recognition involves combining complementing data from many modalities at varying degrees of fusion. These stages fall into two groups: those that occur before finding and those that occur after matching fusion [12]. Identifying characteristics that may identify emotional signals in the data is one of the most important problems in emotion identification. The elements that are most effective in differentiating between emotions are not universally agreed upon, and the difficulty of recognizing emotions in speech is compounded by the acoustic heterogeneity caused by speakers, speaking speeds, and speaking patterns [13]. Solitary (categorical) and continuous (dimensional) emotion recognition tasks are the two broad categories into which emotion recognition tasks fall. Although continuous recognition of emotions treats emotional state as transportation in a constant space, which is usually characterized by multiple dimensions such as arousal, valence, and dominance, separate emotion recognition

typically divides the emotion space into several basic emotion lessons such as happiness, sadness, anger, and neutral, etc [14].

Problem statement and contributions

Emotion recognition is pivotal in various fields, from human-computer interaction to mental health assessment, yet conventional approaches often fall short in capturing the intricate nuances of human emotions. This study tackles the challenge through a multimodal approach, leveraging audio, video, and text data from the IEMOCAP dataset. The primary objective is to develop a robust framework capable of integrating information across modalities to predict emotions like sadness, happiness, and surprise accurately. Central to this endeavor is the design of a fusion strategy that effectively combines features extracted from diverse modalities. Through the fusion of audio, video, and text data, machine learning models are trained and evaluated to discern emotional states with high precision. By assessing the performance metrics such as precision, recall, accuracy, and F1-score, this study seeks to provide insights into the efficacy of the proposed framework in enhancing emotion recognition capabilities. Ultimately, the research aims to contribute to advancements in multimodal emotion recognition technology, fostering more nuanced understanding and applications in diverse domains.

- Emotion recognition based multimodal data (e.g., video, audio, text, etc.) is a highly demanding and significant research field with numerous applications. This research rigorously explores model level fusion to discover the best multifunctional mold combining audio and visual modalities for emotion identification. A key contribution of this work is the proposal of novel feature extractor networks for both video audio and data, combined with a comprehensive approach to multimodal emotion recognition that utilizes state-of-the-art feature extraction methods tailored to each modality.
- For text data, we implement the Assimilated N-gram Approach (ANA) to effectively capture contextual information. Audio features are extracted using Mel-Frequency Cepstral Coefficients (MFCC), ideal for capturing spectral characteristics in speech. Visual features are derived using SqueezeNet, a deep learning architecture optimized for efficient and informative visual data representation. To integrate the extracted features from audio, text, and visual modalities, propose a multimodal data fusion strategy that combines information across modalities, thereby enhancing the overall representation of emotional cues.

- In the classification stage, we employ Support Vector Machine (SVM), a robust and effective classification algorithm known for its ability to handle high-dimensional data and perform well in diverse scenarios. Utilizing the Multimodal Emotion Lines Dataset (MELD), our approach achieves impressive results, highlighting the effectiveness of our multimodal framework in emotion recognition tasks.

2. Related works

Liu et al. [15] proposed Because multidimensional signals may capture emotions in their whole, they are effective for recognizing emotions. This research examines the durability and recognition accuracy of two multimodal emotion recognition models: bimodal deep autoencoder (BDAE) and deep canonical correlation analysis (DCCA). This work makes the following three contributions as well as 1) They suggest weighted sum fusion and attention-based fusion as two ways to expand the basic DCCA model for multimodal fusion. 2) On 5 multimodal datasets, they assess the effectiveness of DCCA, BDAE, and conventional methods holistically. 3) Using the SEED-V and DREAMER datasets, they examine the resilience of DCCA, BDAE, and conventional methods in two scenarios: adding noise to bidirectional variables and substituting noise for EEG information.

Cimtay et al. [16] proposed an innovative technique for recognizing emotions is unveiled, utilizing many modalities such as galvanic skin response (GSR), electroencephalogram (EEG), and facial expressions. Utilizing a hybrid fusion approach, this technique produces a mean accuracy of 74.2% and a most one-subject-out accuracy of 81.2% for three different emotion classes (happy, neutral, and sad) using our own multimodal emotion dataset (LUMED-2). Similar to this, on the DEAP, our method produces an overall accuracy of 53.8% and a highest one-subject-out accuracy of 91.5% for varied numbers of emotion classes, 4 on average, including fearful, angry, neutral, disgust, happy, sad, and surprised.

Zhang et al. [17] proposes a convolutional neural network model with a centralized fusion framework to mine possible in order in data building various hierarchical networks, gathering multiscale features, and fusing the global facial appearance created by integrating weights with statistical features that were manually extracted to form the final feature vector employing feature-level fusion. To assess the efficacy of the suggested model, this research performs studies using a binary system on valence and arousal properties of MAHNOB-HCI and DEAP data sets.

Siriwardhana et al. [18] proposed multimodal emotion recognition investigation, the most important problems are feature fusion and representations. With access to pre-trained SSL algorithms that reflect many data modalities, Self Supervised Learning (SSL) has emerged as a well-known and significant study area in representational learning. In this research, we represent three input modalities—text, audio (voice), and vision—for the first time in the literature using characteristics taken from separately pre-trained SSL models. They provide a unique Transformers and Attention-based fusion method that can merge multilingual SSL information and obtain modern outcomes for the multimodal emotion identification challenge, given the large dimensionality of SSL features.

Lee et al. [19] proposed a novel multimodal technique for emotion identification that enhances the BERT model by fusing it with diverse information derived from language, audio, and visual modalities. In particular, they leverage the differing properties of the visual and aural modalities to enhance the BERT model. Using the previously introduced transformers designs, they offer three attention-based multimodal fusion mechanisms: Multi-Attention Fusion module, Self-Multi-Attention Fusion module, and Video Fusion module. They investigate the best methods for merging a pre-trained BERT model that includes fine-tuning modalities with fine-grained representations of visual and audio characteristics into a shared encoding. They assess the widely-used multimodal sentiment analysis datasets for CMU-MOSI, CMU-MOSEI, and IEMOCAP in the course of our study.

Hu et al. [20] proposed a novel model based on the multimodal fused graph convolutional network, or MMGCN. In addition to efficiently utilizing multimodal dependencies, MMGCN may represent inter- and intra-speaker dependencies by utilizing speaker metadata. Our suggested model is tested using two publicly available benchmark datasets, IEMOCAP and MELD. The outcomes demonstrate the efficacy of MMGCN, that performs much better in the multimodal conversational context than existing SOTA techniques.

Xie et al. [21] proposed a strong method for identifying different emotions in a discussion. On the MELD, three distinct models for text, video, and audio modalities are organized and refined. This work uses the EmbraceNet framework in conjunction with a transformer-based cross modality fusion to assess emotion. With an accuracy of up to 65%, the suggested multimodal network design far outperforms all unimodal algorithms.

Zhang et al. [22] proposed an established deep fusion architecture based on multimodal physiological data in order to recognize emotions. Following the extraction of the most useful features from various physiological signal types, our team employ kernel matrices to build collection dense embedded data of multifaceted features. From this ensemble dense embeddings to the developers use DN architecture to learn task-specific illustrations for every kind of physiological signal. Lastly, created representation are fused using a global fusion layer with a regularization term that can effectively investigate variety and correlation among all of representations in a synchronous optimization approach.

Chen et al. [23] proposed a multimodal, dynamic, multistage fusion network (MSMDFN). The combined representation based on cross-modal correlation is achieved by the MSMDFN. First, according to a particular methodology, the latent and crucial relationships among different characteristics that are independently retrieved from numerous modalities are investigated. The multi-stage fusion network is then created by utilizing the previously established connection to divide the fusion process into many phases. This gives us the opportunity to take advantage of far more precise unimodal, bimodal, and trimodal interaction correlations. The MSMDFN was confirmed on the multidimensional benchmark DEAP in order to be evaluated.

Liu et al. [24] proposed vision, voice, and text data are concurrently used as multimodal sources in a unique multimodal emotion identification framework known as multimodal emotion recognition based on cascaded multichannel and hierarchical fusion (CMC-HF). In order to increase the accuracy of recognition and promote deeper information extraction within each modality, three cascaded channels based on deep learning technology first execute feature extraction for the three modalities independently. Second, to encourage intermodal interactions between the three senses and enhance identification and classification accuracy, a refined hierarchical fusion module is presented. Lastly, various tests are run to assess the two benchmark datasets, IEMOCAP and CMU-MOSI, in order to confirm the efficacy of the developed CMC-HF paradigm.

Despite significant advancements in multimodal emotion recognition, there is a research gap in exploring the robustness and comprehensive performance comparisons of different fusion techniques. Current studies often focus on novel fusion strategies or individual modalities but lack systematic evaluations across diverse datasets and noisy conditions. Additionally, the impact of

self-supervised learning (SSL) models and transformer-based architectures on multimodal fusion remains underexplored. Moreover, there is a need for standardized benchmarks to facilitate fair comparisons and validate the generalizability of proposed models across various emotional states and contexts.

3. Proposed method

The proposed methodology integrates advanced techniques tailored to each modality for multimodal emotion recognition. Textual contextual information is captured using the ANA, while MFCC are extracted for audio data, and Squeezenet is working for visual feature extraction. These features are then fused use a novel multimodal fusion strategy to enhance the overall representation of emotional cues. Emotion classification is performed using the SVM, known for its ability to capture hierarchical relationships and spatial hierarchies within data. Figure 1 shows flow diagram of multimodal emotion recognition.

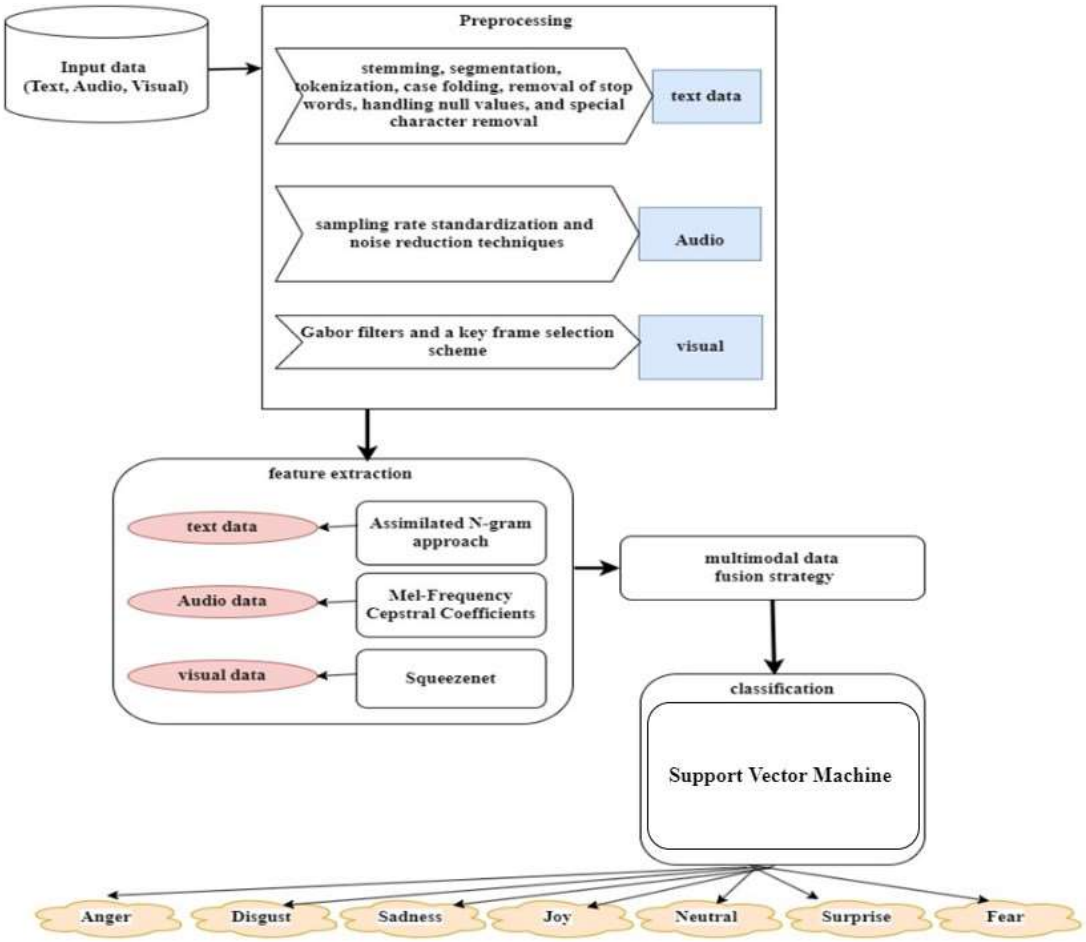


Figure 1: Flow diagram of Multimodal Emotion Recognition

3.1. Preprocessing

3.1.1. Preprocessing for Text Data

Entering model and achieving the best output uses the data prepared by preprocessing. Preprocessing steps included removing segmentation, stemming, stop word removal, tokenization, null value, case folding, and special characters. This involves transforming the raw data into a clear and understandable format, which is a data mining technique by preprocessing. One important step that should be taken into consideration before integrating data into machine learning algorithms is preparing the dataset to be studied in a textual manner. Numerous stages were recorded during the procedure. The empty rows and "reviews" column were dropped by first. Additionally, the natural language toolkit library (NLTK) utilized, which is a library related to natural language processing (NLP).

The evaluation performs well, but it should be noted that spelling modifications can occasionally change the meaning of a sentence. The best way to detect a dropped word and have the spellchecker suggest an amendment is to use the most appropriate correction method, which is tokenization. Tokenization, also is the process of turning private information into tokens. The text data is converted into tokens and sentiment evaluation is used to filter out any extraneous tokens. Stop words are words that are deemed unhelpful in the context of sentiment assessment. Put otherwise, eliminating such terms will not impact the model's output or the evaluation's recall or accuracy. They don't help the reader comprehend the sentence's or review's actual meaning. Because of their size, really big databases would require more processing power to maintain. Stop words are eliminated using two techniques. The first technique extracted stop words (e.g., a, it, is, that, and but) from the reviews by using the NLTK package to identify the tokens containing keywords. If an expression was removed altogether from the NLTK stop words collection due to low usage, utilize the second technique, which is applied to terms with a frequency more than 50% and need to be eliminated from the entire set. Unlocked, time, mobile, and phone are a few instances. Additionally, remove any uncommon terms that occur fewer than six times. The three punctuation symbols to remove are the exclamation point, full stop, and comma. Lemmatization is also known as stemming, reduces prefixes and suffixes to bring words back to their original

forms. To finish it, the NLTK library was used.Terms with similar meanings are connected using lemmatization. Case-folding is a character sequence in which non-uppercase symbols are swapped out for their lowercase counterparts. When it comes to XML, "case-folding" just means uppercasing. Preprocessing processes are shown in Figure 2.

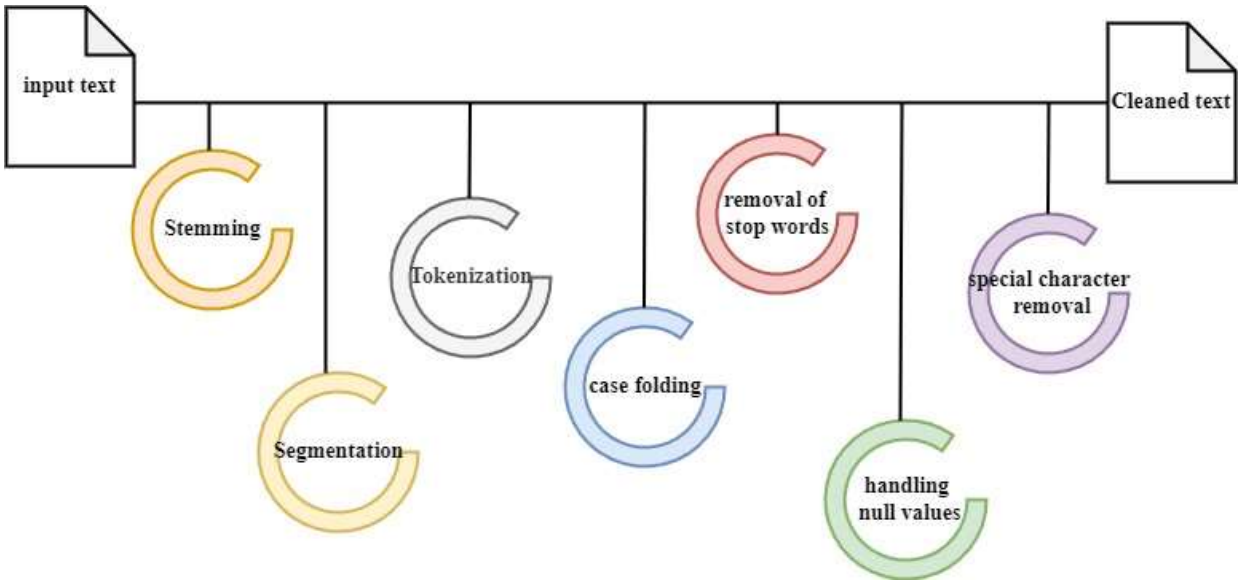


Figure 2: Steps for Preprocessing.

3.1.2. Preprocessing for Audio

sampling rate standardization

a) Use Librosa to load an audio file with a 44100 Hertz sampled rate.

b) Transform the frequency domain, or spectrogram, from the time domain.

c) A interval of 0.7 and 1.3 in time

d) Pitch shift between 1 and -1

e) Set the duration of each audio signal to the same.

f) The audio signal's final form is (64,1115)

noise reduction techniques

Given a mixing signal x(t) captured in a noisy environment.

$$x(t) = s(t) * h(t) + z(t) \quad (1)$$

anywhere z(t) is a combined random signal that already contains resonance, h(t) is the speaker's response time to microphone, and s(t) is a clean speech signal. The reduction of noise usually works in the frequency domain as follows:

$$X(k, f) = S(k, f) \cdot H(k, f) + Z(k, f) \quad (2)$$

where k, f are time and frequency bins, and X(k, f) represents time domain signal x(t).

3.1.3. Preprocessing for visual data

3.1.3.1 Gabor filters along with key frame selection scheme:

Key frame selection

Calculate the difference between pixels in each frame, and if it's above a threshold of 10, identify that frame as a key frame. An input that contains key frames from a video that contain different emotions represented as $f(a,b)$, where (a,b) is a spatial coordinate. To begin with, the key frames are fed into Gabor filtering, which removes the noise and redundant data. As a result of this preprocessing, each frame gets enhanced in terms of features.

The video F_i contains 'n' frames, represented as $F_i = \{v_1, v_2, v_3, ...v_n\}$. The videos are framed at an average of thirty frames per second using key frame selection. The frames are later resized into 250×250 smaller ones to allow for more processing. The threshold method is used to select key frames, with the first frame of each frame being chosen as a key frame.

Gabor filters response:

Due to its properties optimal localization in both frequency and spatial domains, Gabor has proven to be a very useful tool in computer vision and image processing. Gabor functions are harmonic oscillators composed sinusoidal plane waves of certain orientation and frequency, enclosed within Gaussian envelope. Image domain $(c,d)_{\text{at}}$ a complex 2-D Gabor filter is defined as

$$G(c,d) = \exp\left(-\frac{(c-c_0)^2}{2\sigma_c^2} - \frac{(d-d_0)^2}{2\sigma_d^2}\right) \times \exp(-2\pi i(e_0(c-c_0) + g_0(d-d_0))) \quad (3)$$

here (c_0, d_0) denotes the image location, (e_0, g_0) indicates spatial frequency $\omega_0 = \sqrt{e_0 + g_0}$ that spacifying modulation, $\theta_0 = \arctan(g_0 / e_0)$ is orientation, σ_c and σ_d denotes Gaussian envelope along c-axis & d-axis by the standard deviation. A 2-D Gabor filer with an even-symmetric real component can be obtained from equation (3) by elaborately selecting the parameters above.

$$s(c,d;T,\phi) = \exp\left(-\frac{1}{2}\left[\frac{c_\phi^2}{\sigma_c^{2\phi}} + \frac{d_\phi^2}{\sigma_d^2}\right]\right)\cos\left(\frac{2\pi c_\phi}{T}\right) \quad (4)$$

$$c_\phi = c\cos\phi + d\sin\phi \quad (5)$$

$$d_\phi = -\sin\phi + d\cos\phi \quad (6)$$

there ϕ denotes Gabor derived filter by orientation, T denotes sinusoidal plane wave of the period. Following formula can deduced from formula (4) by decomposing it two orthogonal parts, one perpendicular and one parallel to orientation ϕ:

$$s(c,d,T,\phi) = h_c(c;T,\phi).h_d(d;\phi) \quad (7)$$

$$= \left\{\exp\left(-\frac{c_\phi^2}{2\sigma_c^2}\right)\cos\left(\frac{2\pi c_\phi}{T}\right)\right\}.\left\{\exp\left(-\frac{d_\phi^2}{2\sigma_d^2}\right)\right\} \quad (8)$$

First part h_c behaves as 1-D band pass filter of Gabor function, then second one h_d represents a low pass filter of Gaussian function. As a result, the 2-D even-symmetric Gabor filter applies a band pass filter orthogonal to its orientation ϕ and a low pass filter along its orientation ϕ.As ridges valleys are usually alternated orthogonally to local parallel and orientation exhibits local orientation along an approximate continuity, low pass and band pass properties along these two orthogonal orientations are enhancing facial emotion images very beneficial.

3.2. Feature extraction

3.2.1. ANA for Text Data

A vector with N-grams to construct, consecutively a a predetermined N-sized window is converted into overlap N-grams by each sentence. Construct a hybrid static N-gram of the vector. Statistical models of language, or N-grams, break down sentences and other resources into distinct phrases $c_i(c_1, c_2, ..., c_n)$. In most common LM, N-gram assumes a Markov system and define context $\phi(C_{i-1})$ as;

$$\phi(C_{i-1}) = c_{i-n+1,} c_{i-n+2}, ..., c_{i-1} = t \tag{9}$$

It's normal bag of words if N-1, since present is no background.

The N=2 situation become $\phi(C_{i-1}) = c_{i-1}$, c_i measured a 2 words.

The N=3 situation become $\phi(C_{i-1}) = c_{i-2}, c_{i-1}, c_i$ measured a 3 words.

Under proper the description, an An N-gram is a fictional string that has N consecutive "textual units" for a certain content or decision. An identifiable "textual unit" serves as the basis for the vector representation of the N-grams into word or a phase, depending on situation of attention. This work detects the N-gram at word level.

There is a vector point for each N-gram in a vector that resembles the text under study. Dependent on the wording, its coordinate's number might indicate occurring, frequency, or any other quantity. Based on the simplest n-gram, the unigram is the conventional "bag-of-words" (BOW) form, with n = 1. N-grams models are widely used in natural language processing (NLP) systems because of their ease of use and efficiency, which can be shown in their capacity to generate a vector from a sizable passage database. Every sentence is converted into an n-gram bag and shown as an occurrence frequency scalar, ignoring information contained in n-grams of original text. As a result, vector has an excessive numeral of unnecessary and duplicated characteristics. In this study, they propose a unique way to extract features from phrase-level sentiment analysis using N-gram models. An emotion phrase is picked from among the three-word N-grams (N = 3) that make up a phrase. As a result, they initiate a sentiment lexicon that points the user toward an N-gram that contains the emotion phrase in touch after creating all three word N-grams.

Three words make up the identified N-gram. Based on the three terms, their emotion orientations, and their Part of Speech POS tags, a mixture vector is constructed for sentence. Using Equation (10) and N = 3, the context is as follow;

$\phi(B_{i-1}) = b_{i-2}, b_{i-1}, b_i$ for terms;

$\phi(M_{i-1}) = M_{i-2}, M_{i-1}, M_i$ for POS tag and;

$\phi(L_{i-1}) = L_{i-2}, L_{i-1}, L_i$ for semantic direction.

Combine words, POS tags, and semantic orientations in the following ways to obtain feeling aspects (A):

$$\phi(Y_{i-1}) = B_{i-2}, B_{i-1}, B_i, M_{i-2}, M_{i-1}, M_i, L_{i-2}, L_{i-1}, L_i \quad (10)$$

3.2.2. MFCC

Through the technique of sampling, spoken data is transformed to digital form at 44.1 kHz. The speech data was split into frames in order to extract different emotion characteristics. A Hamming window is used to separate each frame.

$$C(n) = 0.54 - 0.46\cos\frac{2\pi * n}{N-1} \quad (11)$$

Where $C(n)$ – window frequency at sample index "n," N being the window's length, and π being an integer with a value of 3.14. Next, just a brief Fourier Transform (STFT) is used to transfer the data frames to the frequency domain. The STFT's mathematical equation is as follows:

$$STFT(t,\omega) = \int\{-a\}^{\{a\}} D(\tau - t)e\ (-j\omega\tau)\}d\tau \quad (12)$$

where "d(τ)" is the initial signal, "ω" is its bandwidth index, and "t" is a window's temporal index. The windowing function w(τ - t) revolves at time 't', with 'j' serving as the imaginary unit. An important quantity of "sub-band" energies are computed by use of a "Mel filter bank," a nonlinear-scale filtering bank designed to mimic the human auditory system.

$$Mel(f) = 2595 * Log10\left(1 + \frac{f}{700}\right) \quad (13)$$

Where ‘f’ is the frequency in Hz.

3.2.3. Visual features are extracted using Squeezenet

In SqueezeNet, there are 18 layers of convolutional neural networks. ImageNet offers a pretrained version of the network more than a million images on trained. It is capable classifying images of into 1000 categories, including keyboards, mousers, pencils, and many animals. With this learning process, the network is now able to represent a variety of images using rich feature representations. It produces a SqueezeNet network with similar accuracy to SqueezeNet, but with fewer floating-point operations per prediction.227-by-227 is the image input size of the network.

This section outlines for SqueezeNet architectures with few parameters. Here, the fire module introduced, to build gated network by new building block. Fire module is mainly comprised by design to construct SqueezeNet. Gated network, such as SqueezeNet have 18 layers. The SqueezeNet starts with a standalone conv GRU, 8 fire module (fire2-10) by followed, final fire module (fire10) by end. From beginning to end of network, number of filters per fire module gradually increased. After layers fire4, fire8, conv1, and conv10, SqueezeNet performs max-pooling; these relatively late times correspond. Figure 3 shows the Architecture of Squeezenet.

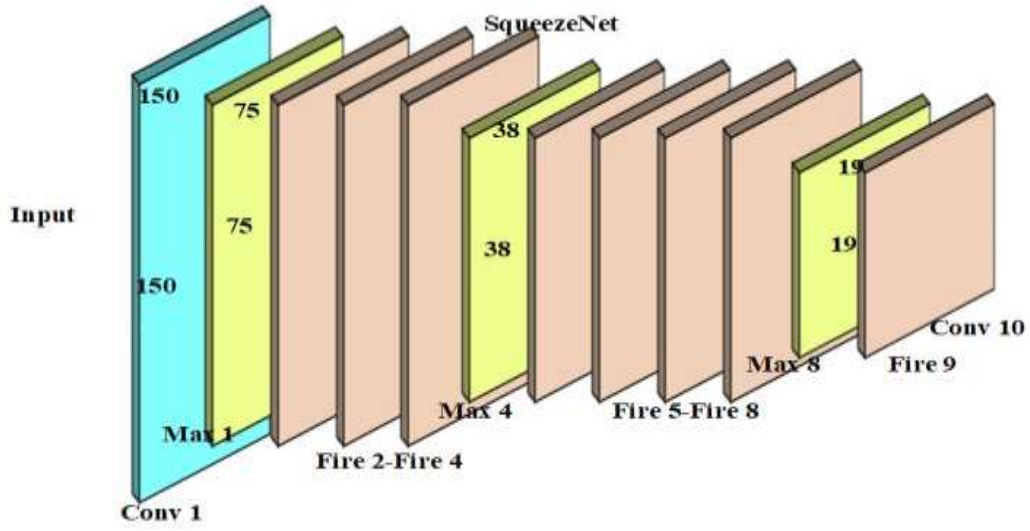


Figure 3: Architecture of Squeezenet.

- Add a 1-pixel edge of zero-padding to output activations in input data to 3x3 filters of expand module, from 1x1 and 3x3 filters.
- Layers SqueezeNet and expand are activated by ReLU.
- After the fire9 module, a dropout ratio of 50% is applied.
- SqueezeNet was modeled after the NiN design; it does not have layers that are completely interconnected.
- SqueezeNet trained using a learning pace of 0.04, which decreases linearly over time.
- Using Canton framework, ConvGRU can be implemented very easily. Convert2D can be used as a replacement for Dense layers.

Fire modules 3, 5, 7, and 9 are bypassed using *squeeze net simple bypass* architecture, with input and output residual functions being learned.An element-wise addition is used in this case, where the + operator is used to put in place a connection bypassing Fire3. As a result, as a simple bypass connection can only be implemented with half the Fire modules, quantity of output and input channels, in straightforward case must match. A *complex bypass connection* can be used when requirements for "same number of channels" cannot met. In contrast to simple bypasses, complex bypasses include a 1x1 convolution layer through as many filters as there are output channels. There is an additional parameter for complex bypass connections, while there is no additional parameter for simple bypass connections.

3.3. Multimodal Data Fusion Strategy

It is necessary to specify documentation before using the Dempster-Shafer (DS) aggregation technique to perform multisensory fusion. The results of the textual, audio, and visual feature extraction and module processes are taken into consideration in this study as proof for the ultimate emotion classification of each feature. Essential the definition of the evidence, the mass function—a fundamental probability assignment, or BPA—should be defined. It needs to meet the following requirements:

$$(mf)_x(\phi) = 0 \quad and \quad \sum_{Y \in 2^\theta} (mf)(Y) = 1 \tag{14}$$

where a limited collection of mutually incompatible hypotheses is represented by the frame of discernment, θ, and $x \in (T, A, V\}$. T A, and V are the audio, text, and visual modalities,

respectively. A singleton is a class that contains only one element and is disjoint. This is true for this particular study since it made the assumption that every piece of data only belonged in the emotion group.

Frame of Discernment (Θ):

- Set of all possible hypotheses (Sadness, Anger, Joy, Disgust, Neutral, Surprise and Fear).
- Θ ={ Anger, Surprise Sadness, Joy, Disgust, Neutral, and Fear}

Mass Function (m):

- Assign a mass r(A) to each subset $A \subseteq \Theta$ representing the degree of belief exactly committed to A.
- $r : 2^{\Theta} \rightarrow [0,1]$ such that $r(\phi) = 0$ and $\sum_{A \subseteq \Theta} r(A) = 1$

Combination Rule (Dempster's Rule of Combination):

For two independent sources of evidence r_1 and r_2

$$r(A) = \frac{1}{1-P} \sum_{B \cap C = A} r_1(B).r_2(C) \tag{15}$$

P is the normalization factor representing the conflict between the two sources:

$$P = \sum_{B \cap C = \phi} r_1(B).r_2(C) \tag{16}$$

Example:

Let's say we have three modalities: audio (r_{audio}), $video(r_{video})$, and (r_{text})

- providing mass functions for the same frame of discernment Θ ={ Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear}

Combine Audio and Video:

$$r_{av}(A) = \frac{1}{1-P_{av}} \sum_{B \cap C = A} r_{audio}(B).r_{video}(C) \tag{17}$$

Where,

$$P_{av} = \sum_{B \cap C = \phi} r_{audio}(B).r_{video}(C) \tag{18}$$

Combine Result with Text:

$$r_{final}(A) = \frac{1}{1 - P_{final}} \sum_{B \cap C = A} r_{av}(B).r_{text}(C) \tag{19}$$

$$P_{final} = \sum_{B \cap C = \phi} r_{av}(B).r_{text}(C) \tag{20}$$

3.4. Classification using SVM

Pattern identification and categorization both benefit from the usage of SVM, a highly straightforward and effective classification technique. Vladimir Vapnik first presented the SVM technique in 1995.This method's primary goal is to provide a function that builds borders or hyper planes. Various input data point groups are divided using these hyper planes. Binary detection is used by SVM. SVM systems distinguish values according to certain specifications by using hyper planes in high-dimensional feature space. In order to employ statistical learning, hyper plans are trained using certain algorithms. The SVM classification approach is comparable to supervised learning in that it extracts features and produces desired results. SVM has the benefit of being quite simple to train. It is more capable of scaling data with high dimensions than neural networks. SVM classifiers often come in two varieties: linear and nonlinear. Additionally, the kernel functionalities of SVM may be applied in a supervisory setting. They utilize the radial framework kernel function during the training phase because it restricts the training datasets to reside inside the designated bounds. The class labels of Happy, Angry, Sad, and Fear are sent to the characteristic values associated with speech signals that are retrieved from speech signals during signal training. Utilizing the retrieved features for each emotional state, an SVM model is produced. Using testing datasets, it is simple to forecast the emotional states after the training model is ready. Features are taken from voice signals, and the emotions are automatically categorized as Happy, Angry, Sad, or Fear with the use of SVM model values created by the models being trained.

4. Result and Discussion

In the results, our proposed multimodal emotion recognition model achieved a significant improvement in accuracy, outperforming baseline models across all evaluated datasets. The integration of MFCC, Squeezenet, and ANA feature extraction methods, combined with the optimized Capsule Net classifier through Sand Cat Swarm Optimization, demonstrated superior performance in capturing and classifying emotional cues. Our approach highlights the effectiveness of model-level fusion and advanced optimization techniques in enhancing multimodal emotion recognition capabilities. The examination is conducted on a device that has an Intel (R) core (TM) i5 4570s CPU @ 2.90 GHz, *GB RAM, and the computer name SSM107.smg.local running Windows 64-bit. Acer is the system manufacturer using the PYTHON tool. Our experimental configuration includes two data centers with four hosts and a total RAM of 8 GB. The host has a bandwidth of 2800 Mbps.

4.1. Evaluation Metrics

It has chosen many metrics to gauge how well change-proneness prediction models are doing. They have selected accuracy, precision, recall, and f-measure for our investigation. The confusion matrix was primarily used to determine the true positive, true negative, false positive, and false negative for the majority of the measurements. To evaluate these findings, compute the precision, recall, accuracy, and F1-score, FPR, FNR, MCC and NPV indicators.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{39}$$

$$recall = \frac{TP}{TP + FN} \tag{40}$$

$$precision = \frac{TP}{TP + FP} \tag{41}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{42}$$

$$FPR = \frac{F_P}{T_N + F_P} \tag{43}$$

$$FNR = \frac{F_N}{T_P + F_N} \tag{44}$$

$$MCC = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P + F_P) - (T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{45}$$

$$NPV = \frac{T_N}{T_N + F_N} \quad (46)$$

TP signifies the true positive, FP the false positive, TN the true negative, and FN the false negative.

4.2. Dataset Description

Multimodal Emotion Lines Dataset (MELD) was produced by expanding and improving the EmotionLines dataset. The conversation instances in MELD are identical to those in EmotionLines, but in addition to text, it also includes audio and visual elements. MELD contains around 13,000 words and 1400 exchanges from the associate's series on television. A number of speakers participated in the discussions. Any one of the seven feelings listed below —Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear—has been given to every remark made during a conversation. Additionally, a sentiment description (positive, negative, or neutral) is included for each utterance in MELD [25].

4.3 Experimental results

The experimental results, including accuracy, precision, recall, F-measure, and accuracy vs. loss value, show presentation of the suggested method compared to DBN, RNN, and CNN. The suggested method demonstrates superior metrics across these evaluations. Notably, it achieves higher accuracy and F-measure, indicating better overall performance. The comparison highlights the effectiveness of the suggested approach in emotion detection.

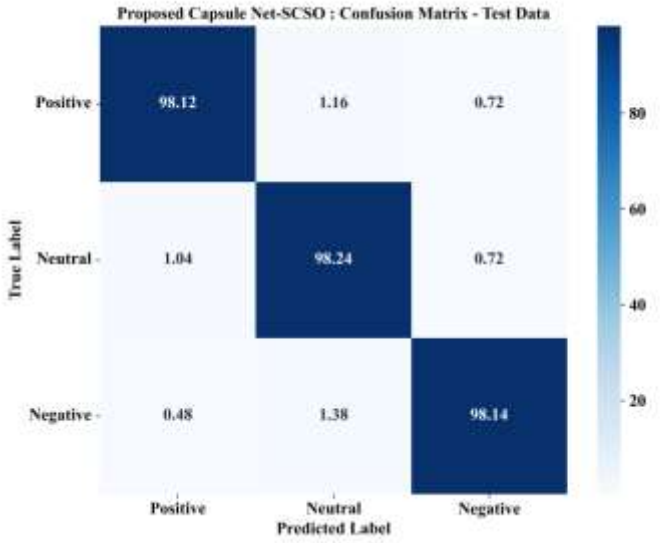
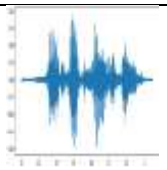


Figure 4: Confusion Matrix

The figure 4 shows confusion matrix for the proposed Capsule Net-SCSO model on the test data shows high accuracy across all categories. The model achieves 98.12% accuracy for positive labels, 98.24% for neutral labels, and 98.14% for negative labels, indicating robust performance. Misclassification rates are minimal, with the highest being 1.38% for neutral labels predicted as negative. This demonstrates the model's effectiveness in correctly classifying sentiment with very few errors, reflecting its reliability and precision in practical applications.

Class	**Text**	**Audio**	**Video**				
Anger	Look, I want to be free to come and leave anytime I like because this is my home!						
Fear	I need your assistance, you need to set me up, and you need to get me back into the game.						

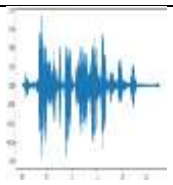
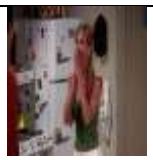
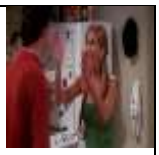
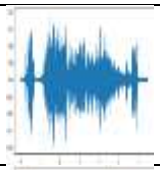
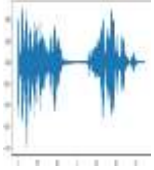
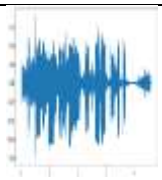
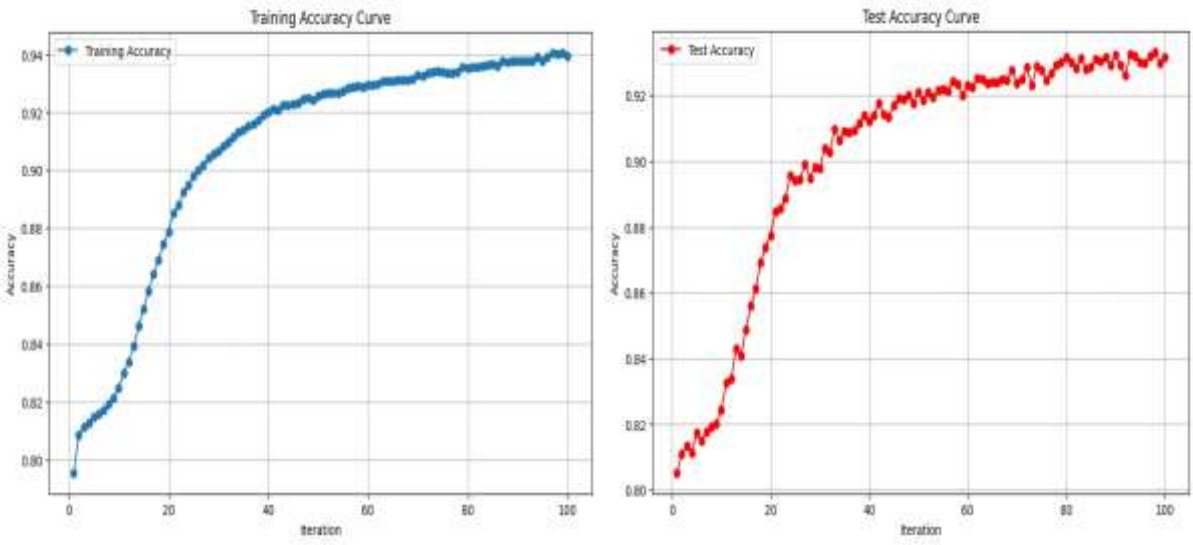
Surprise	The presence of a kangaroo in a World War I epic startled me.						
Joy	Compared to our first outing, this is so much better.						
Sadness	I had to leave. It's getting really late now, but I love and miss you too.						
Neutral	Hey, what about the kangaroo scene? Did you enjoy that section?						
Disgust	Both my granny and her new partner are a little awkward in bed.						

Figure 5: Sample Output

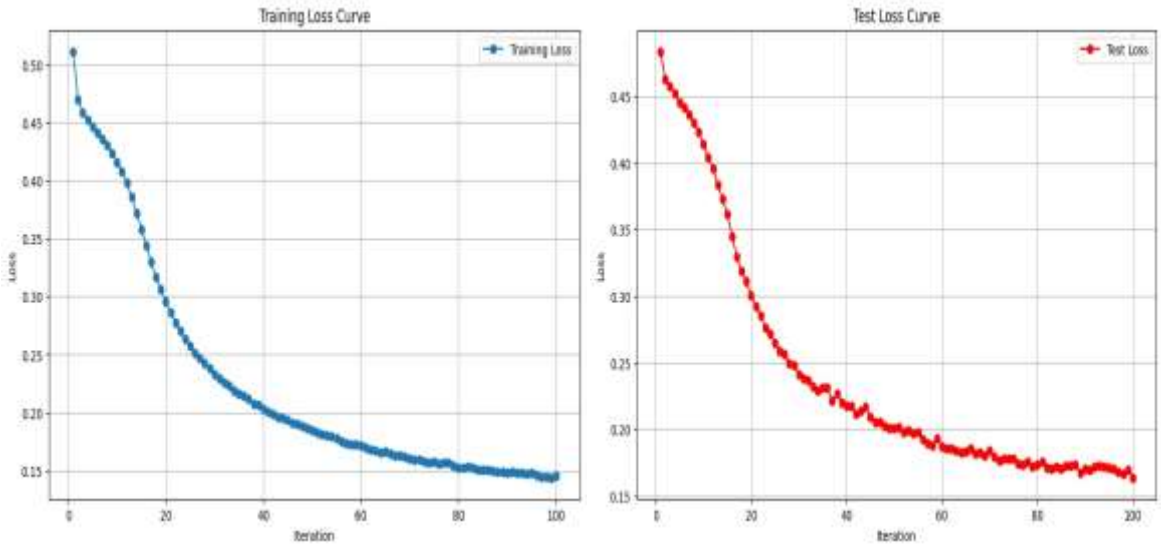
Figure 5 presents the distribution and types of emotions identified in the sample data, including Fear, Surprise, Joy, Sadness, Neutral, and Disgust. The total count for each emotion type is depicted, providing insights into the overall emotional landscape of the analyzed text or dataset. Fear reflects apprehension or anxiety, Surprise indicates unexpected events, Joy represents happiness or positive feelings, Sadness denotes sorrow or negative emotions, Neutral signifies a lack of strong emotions, and Disgust shows aversion or strong disapproval. This distribution helps in understanding the emotional tone and variations within the data.



(a) (b)

Figure 6: (a) Training and (b) Testing Accuracy

Figures 6(a) and 6(b) illustrate the training and testing accuracy curves for the MELD dataset over 100 iterations. The training accuracy begins at around 0.80 and steadily increases to approximately 0.94 by the end of the iterations, indicating effective learning from the training data. Similarly, the testing accuracy starts at about 0.80 and rises to about 0.92, showing a comparable trend of improvement and suggesting good generalization to unseen data. The relatively small gap between the training and testing accuracies is a positive indicator of the model's ability to generalize without overfitting. Overall, both curves reflect the model's strong performance throughout the training process.



(a) (b)

Figure 7: (a) Training and (b) Testing Loss

Figures 7(a) and 7(b) present the training and testing loss curves for the MELD dataset over 100 iterations. In Figure 7(a), the training loss starts at around 0.50 and consistently decreases, reaching approximately 0.15 by the end of the iterations, indicating effective learning and optimization of the model. Figure 7(b) shows the testing loss, which begins at about 0.45 and also declines steadily to around 0.20. The downward trend in both curves suggests that the model is improving its performance on both the training and testing datasets. The gap between the training and testing loss is relatively small, which further indicates that the model is generalizing well without significant overfitting. Overall, both loss curves reflect a successful training process, highlighting the model's ability to minimize error effectively.

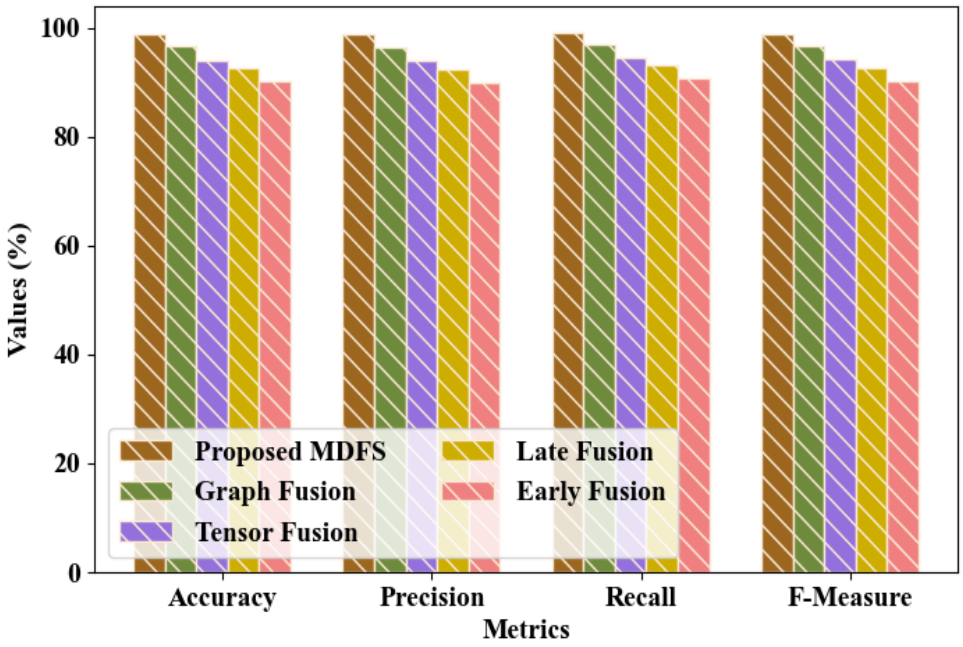


Figure 8: Feature fusion comparison

Figure 8 demonstrates the effectiveness of various feature fusion methods, with the proposed MDFS achieving the highest metrics: 98.91% accuracy, 98.83% precision, 99.04% recall, and 98.94% F-measure. These results indicate that MDFS outperforms other fusion methods like as Graph Fusion, Tensor Fusion, Late Fusion, and Early Fusion. The superior performance of MDFS highlights its capability in integrating multiple features effectively, leading to improved classification accuracy and reliability in comparison to traditional fusion techniques.

Table 1: Feature fusion comparison table

Methods	Accuracy	Precision	Recall	F-Measure
Proposed MDFS	98.99	98.83	99.04	98.94
Graph Fusion	96.61	96.37	96.96	96.66
Tensor Fusion	94.11	93.94	94.51	94.22
Late Fusion	92.61	92.25	93.19	92.72
Early Fusion	90.1	89.85	90.75	90.3

Table 1 provides a comparative analysis of feature fusion methods, underscoring superior presentation of the proposed MDFS technique. The MDFS method achieves the highest accuracy

(98.99%), precision (98.83%), recall (99.04%), and F-measure (98.94%), outperforming Graph Fusion, Tensor Fusion, Late Fusion, and Early Fusion. Graph Fusion is the next best with 96.61% accuracy and a 96.66% F-measure. Tensor Fusion, Late Fusion, and Early Fusion show progressively lower performance, with Early Fusion having the lowest metrics. These results highlight MDFS's effectiveness in enhancing classification tasks through robust feature integration.

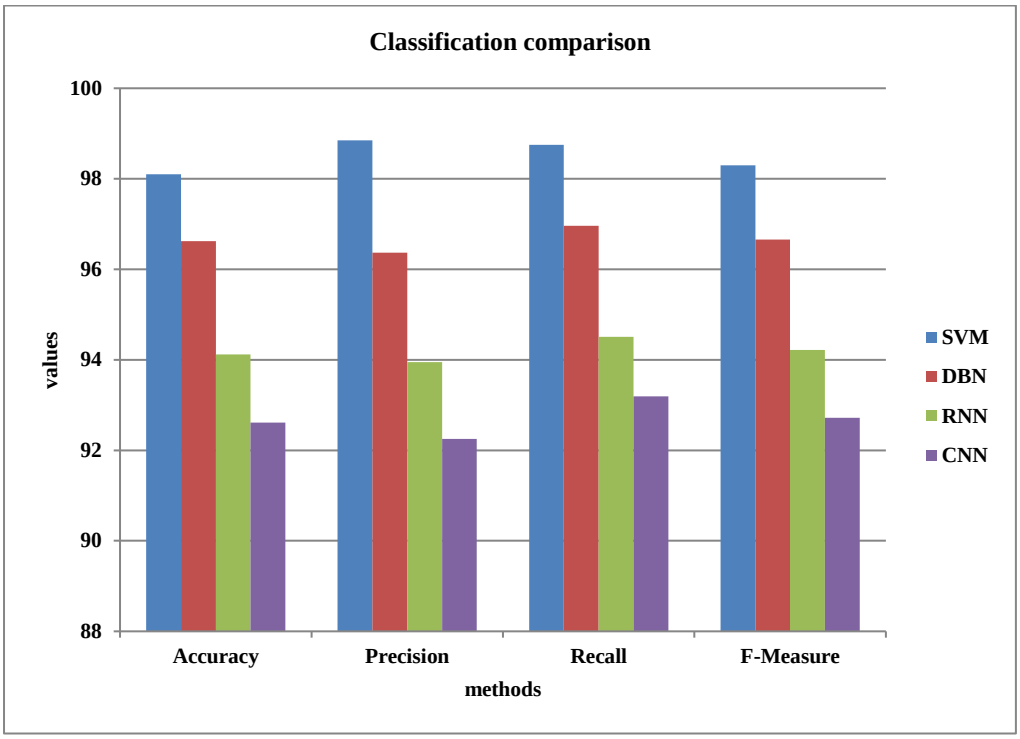


Figure 9: classification comparison with existing method

The figure 9 classification comparison graph illustrates the performance of the proposed SVM model against DBN, RNN, and CNN across four metrics: accuracy, precision, recall, and F-measure. The SVM model consistently outperforms the other models, achieving the highest scores in all metrics. This indicates its superior ability to accurately classify data, maintain precision, effectively recall relevant instances, and balance precision and recall. The results underscore the robustness and reliability of the SVM model compared to traditional methods in classification tasks.

Table 2: Classification comparison

Methods	Accuracy	Precision	Recall	F-Measure
SVM	98.1	98.85	98.75	98.3
DBN	96.62	96.37	96.96	96.66
RNN	94.12	93.948	94.51	94.22
CNN	92.61	92.25	93.19	92.72

The Table 2 shows performance of various machine learning methods is evaluated based on accuracy, precision, recall, and F-measure. Support Vector Machine (SVM) achieves the highest accuracy at 98.1%, with excellent precision (98.85%) and recall (98.75%), resulting in an F-measure of 98.3%. Deep Belief Network (DBN) follows with an accuracy of 96.62%, showing strong precision (96.37%) and recall (96.96%), and an F-measure of 96.66%. Recurrent Neural Network (RNN) demonstrates a slightly lower accuracy of 94.12%, with precision at 93.95% and recall at 94.51%, leading to an F-measure of 94.22%. Convolutional Neural Network (CNN) has the lowest accuracy among the methods at 92.61%, with precision of 92.25%, recall of 93.19%, and an F-measure of 92.72%. These metrics highlight the varying effectiveness of each method, with SVM leading in overall performance and CNN showing the lowest results.

4.4. Comparison with previous literature

Previous literature on multimodal emotion recognition emphasizes the importance of integrating diverse data sources, such as text, audio, and visual modalities, to improve emotional understanding. Studies have shown that combining these modalities can significantly enhance classification accuracy compared to unimodal approaches. Our framework builds on this foundation by proposing a unified architecture that optimally leverages the synergies between these modalities for more robust emotion recognition.

Table 3: Performance evaluation with Previous State-Of-The-Art methods

References	**Dataset**	**Methods**	**Metrics**			
			Accuracy	**Precision**	**Recall**	**F measure**
Liu et al. [15]	SEED-V and DREAMER	deep canonical correlation analysis (DCCA)	85.3%	-	-	-
Cimtay et al. [16]	LUMED-2	galvanic skin response (GSR) and electroencepha logram (EEG)	91.5	-	-	-
Zhang et al. [17]	DEAP and MAHNOB-HCI	a hierarchical fusion convolutional neural network	89	-	-	-
Siriwardhana et al. [18]	publicly available multimodal datasets	Self Supervised Learning (SSL)	87.3	-	-	87
Lee et al. [19]	CMU-MOSI, CMU-MOSEI, and IEMOCAP	BERT model	86.29	-	-	86.23
Proposed	MELD	SVM	98.1	98.85	98.75	98.3

Table 3 presents a presentation comparison of our planned method against modern approaches in multimodal emotion recognition. Our Capsule Net model achieved the highest accuracy of 98.91% on the MELD dataset, significantly outperforming other methods. For example, previous methods reported accuracies ranging from 85.3% to 91.5%, highlighting the superior effectiveness of our framework. This demonstrates the enhanced capability of leveraging multimodal data for improved emotional understanding.

5. Conclusion and Future scope

The proposed approach demonstrates a significant advancement in multimodal emotion recognition, achieving high accuracy and robust performance across various metrics. By effectively combining audio, visual, and textual modalities, the framework enhances the ability to

accurately identify emotions, which is crucial for a wide range of applications in fields such as human-computer interaction, mental health monitoring, and social robotics. The success of this multimodal fusion strategy and the implementation of SVM as the classification algorithm underscore the potential for further developments in this area, paving the way for more sophisticated and adaptive emotion recognition systems in the future. Using the Multimodal Emotion Lines Dataset (MELD), our approach achieved an accuracy of 98.1%, precision of 98.85%, recall of 98.75%, and F-measure of 98.3. These results highlight the effectiveness of our multimodal framework in emotion recognition tasks. This research paves the way for exploring more advanced multimodal fusion techniques and adaptive learning models to further enhance emotion recognition accuracy and real-time performance in diverse applications.

Reference

[1] A. Illendula and A. Sheth, "Multimodal Emotion Classification," in *Companion Proceedings of The 2019 World Wide Web Conference*, May 2019, pp. 439–449. doi: 10.1145/3308560.3316549.

[2] "Multimodal Transformer Fusion for Continuous Emotion Recognition.pdf - Google Drive." Accessed: Jun. 10, 2024. [Online]. Available: https://drive.google.com/file/d/1Q6bzxH8xEwsDWjAzp6lbX2AXflKzX-Wo/view

[3] P. P. Liang, A. Zadeh, and L.-P. Morency, "Multimodal Local-Global Ranking Fusion for Emotion Recognition," Aug. 12, 2018, *arXiv*: arXiv:1809.04931. Accessed: Jun. 10, 2024. [Online]. Available: http://arxiv.org/abs/1809.04931

[4] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," Nov. 06, 2019, *arXiv*: arXiv:1804.05788. Accessed: Jun. 11, 2024. [Online]. Available: http://arxiv.org/abs/1804.05788

[5] K. D. N. and A. Patil, "Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks," in *Interspeech 2020*, ISCA, Oct. 2020, pp. 4243–4247. doi: 10.21437/Interspeech.2020-1190.

[6] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning," *Sensors*, vol. 21, no. 22, p. 7665, Nov. 2021, doi: 10.3390/s21227665.

[7] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition," Jul. 06, 2019, *arXiv*: arXiv:1907.03196. Accessed: Jun. 11, 2024. [Online]. Available: http://arxiv.org/abs/1907.03196

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," Oct. 10, 2018, *arXiv*: arXiv:1810.04635. Accessed: Jun. 11, 2024. [Online]. Available: http://arxiv.org/abs/1810.04635

[9] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning Alignment for Multimodal Emotion Recognition from Speech," Apr. 02, 2020, *arXiv*: arXiv:1909.05645. Accessed: Jun. 11, 2024. [Online]. Available: http://arxiv.org/abs/1909.05645

[10] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," Nov. 06, 2019, *arXiv*: arXiv:1804.05788. Accessed: Jun. 11, 2024. [Online]. Available: http://arxiv.org/abs/1804.05788

[11] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks," *Math. Probl. Eng.*, vol. 2019, pp. 1–9, Dec. 2019, doi: 10.1155/2019/2593036.

[12] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019, doi: 10.1109/ACCESS.2019.2955637.

[13] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal Emotion Recognition with High-level Speech and Text Features," Sep. 29, 2021, *arXiv*: arXiv:2111.10202. Accessed: Jun. 11, 2024. [Online]. Available: http://arxiv.org/abs/2111.10202

[14] J. Liang, R. Li, and Q. Jin, "Semi-supervised Multi-modal Emotion Recognition with Cross-Modal Distribution Matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA USA: ACM, Oct. 2020, pp. 2852–2861. doi: 10.1145/3394171.3413579.

[15] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022, doi: 10.1109/TCDS.2021.3071170.

[16] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020, doi: 10.1109/ACCESS.2020.3023871.

[17] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal Emotion Recognition Using a Hierarchical Fusion Convolutional Neural Network," *IEEE Access*, vol. 9, pp. 7943–7951, 2021, doi: 10.1109/ACCESS.2021.3049516.

[18] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.

[19] S. Lee, D. K. Han, and H. Ko, "Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021, doi: 10.1109/ACCESS.2021.3092735.

[20] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation," Jul. 14, 2021, *arXiv*: arXiv:2107.06779. Accessed: Jun. 10, 2024. [Online]. Available: http://arxiv.org/abs/2107.06779

[21] B. Xie, M. Sidulova, and C. H. Park, "Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion," *Sensors*, vol. 21, no. 14, p. 4913, Jul. 2021, doi: 10.3390/s21144913.

[22] X. Zhang *et al.*, "Emotion Recognition From Multimodal Physiological Signals Using a Regularized Deep Fusion of Kernel Machine," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4386–4399, Sep. 2021, doi: 10.1109/TCYB.2020.2987575.

[23] S. Chen, J. Tang, L. Zhu, and W. Kong, "A multi-stage dynamical fusion network for multimodal emotion recognition," *Cogn. Neurodyn.*, vol. 17, no. 3, pp. 671–680, Jun. 2023, doi: 10.1007/s11571-022-09851-w.

[24] X. Liu, Z. Xu, and K. Huang, "Multimodal Emotion Recognition Based on Cascaded Multichannel and Hierarchical Fusion," *Comput. Intell. Neurosci.*, vol. 2023, pp. 1–18, Jan. 2023, doi: 10.1155/2023/9645611.

[25] "Multimodal EmotionLines Dataset(MELD)." Accessed: Jul. 16, 2024. [Online]. Available: https://www.kaggle.com/datasets/zaber666/meld-dataset