Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

A Unified Framework for Multimodal Emotion Recognition: Leveraging Text, Audio, and Visual Data for Enhanced Emotional Understanding

Sanjeeva Rao Sanku[1] and 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

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract:

Emotion recognition based on 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 find the best multifunctional model combining audio and visual modalities for emotion identification. Specifically, it proposes novel feature extractor networks for both audio and video 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, propose a multimodal data fusion strategy that combines information across modalities, thereby enhancing the overall representation of emotional cues. In the classification stage, employ Capsule Net, a novel neural network architecture adept at capturing hierarchical relationships and spatial hierarchies within data, making it well-suited for handling complex multimodal data. To further optimize the performance of the Capsule Net classifier, utilize hyper parameter tuning through the Sand Cat Swarm Optimization (SCSO) algorithm. SCSO, a metaheuristic optimization technique inspired by the behavior of sand cats, iteratively updates candidate solutions to converge towards optimal hyperparameter configurations. Using the Multimodal Emotion Lines Dataset (MELD), our approach achieved an accuracy of 98.91%, precision of 98.83%, recall of 99.04%, and F-measure of 98.94. These results highlight the effectiveness of our multimodal framework in emotion recognition tasks.

Keywords: Assimilated N-gram Approach, Mel-Frequency Cepstral Coefficients, Squeezenet, Capsule Net, Sand Cat Swarm Optimization

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 accuracy, precision, recall, 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.

- For text data, we implement stemming, tokenization, removal of stop words, segmentation, case folding, handling null value, and special character removal. Audio preprocessing involves sampling rate standardization and noise reduction techniques, while visual data undergoes preprocessing with Gabor filters and a key frame selection scheme. This meticulous preprocessing ensures that input data is clean, standardized, and ready for feature extraction.
- We employ state-of-the-art feature extraction methods tailored to each modality. For text data, we utilize the ANA to capture contextual information effectively. Audio features are extracted using MFCC, which are well-suited for capturing spectral characteristics in speech. Visual features are extracted using Squeezenet, a deep learning architecture optimized for efficient yet informative representation of visual data.
- Leveraging the extracted features from text, audio, and visual modalities, we propose a multimodal data fusion strategy to combine information across modalities effectively. This fusion approach enables the integration of complementary information from different modalities, enhancing the overall representation of emotional cues.
- In the classification stage, we employ Capsule Net, a novel neural network architecture known for its ability to capture hierarchical relationships and spatial hierarchies within data. By utilizing Capsule Net for emotion classification, we aim to leverage its inherent advantages in handling complex multimodal data.
- To optimize the performance of the Capsule Net classifier, employ hyperparameter tuning using the SCSO. SCSO is a metaheuristic optimization technique inspired by the performance of sand cats, which iteratively updates candidate solutions to converge towards optimal hyperparameter configurations. This approach ensures that the Capsule Net classifier is fine-tuned to achieve the best possible performance on the given emotion recognition task.

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 electroencephalogram (EEG), galvanic skin response (GSR), 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 the possible information in data by 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 the valence and arousal properties of the 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: the Self-Multi-Attention Fusion module, 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 audio and visual 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 these ensemble dense embeddings to the developers use a deep network 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 the variety and correlation among all of the 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 Capsule Net architecture, known for its ability to capture hierarchical relationships and spatial hierarchies within data. Furthermore, hyperparameter tuning is conducted using the SCSO algorithm, iteratively refining candidate solutions to achieve optimal configurations. Through this comprehensive approach, we aim to significantly

improve the accuracy and robustness of multimodal emotion recognition. Figure 1 shows flow diagram of multimodal emotion recognition.

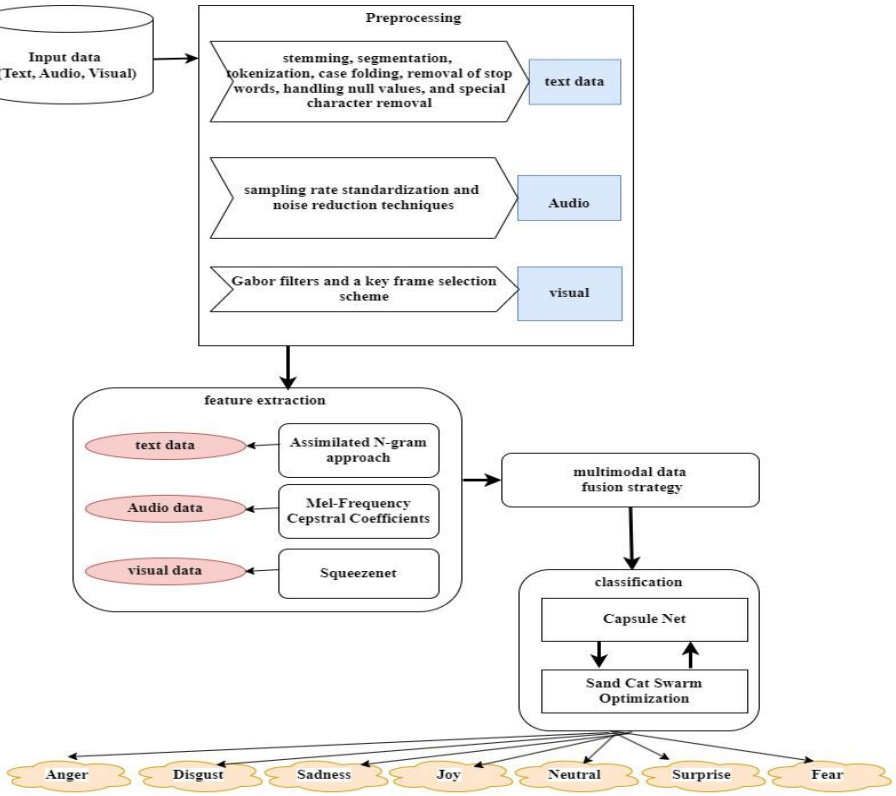


Figure 1: Flow diagram of Multimodal Emotion Recognition

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.

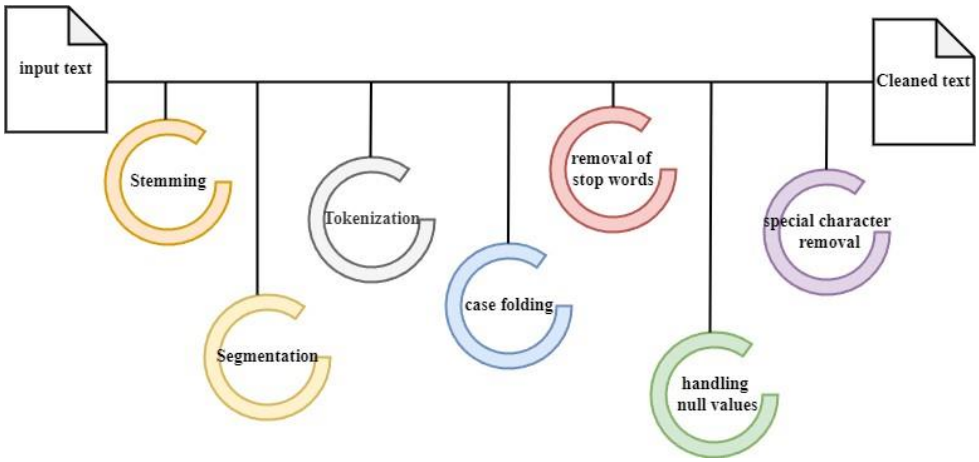


Figure 2: **Steps for Preprocessing.**

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.

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

where z(t) is a combined random signal that already contains resonance, h(t) is the speaker's response time to the 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 the time and frequency bins, and X(k, f) represents the 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) 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 using N-grams to construct, running a a predetermined N-sized window is converted into overlap N-grams by each sentence. Build a hybrid static N-gram of vector. N-grams are statistical language models (LMs) that decompose phrases and materials into individual words $c_i(c_1, c_2, ..., c_n)$. In most common LM, N-gram assumes a Markov system and define context $\phi(C_{i-1})$ as;

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

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

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

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

Under proper the description, an N-gram is narrative string including N contiguous "textual units" as of a given ruling or material. A "textual unit" that may be recognized as a character forms a vector illustration of the N-grams into word or a phase, depending on situation of attention. This work detects the N-gram at word level.

Every N-gram in a vector that resembles the text being studied corresponds to a vector position. That coordinate's value may represent the frequency, happening, or any other measure, depending on the text. The unigram is the standard "bag-of-words" (BOW) form, with n = 1, according to the most basic n-gram. N-grams models be frequently utilized in NLP applications due to their simplicity and effectiveness, as demonstrated by their ability to produce a vector from a large passage database. Each phrase is transformed into a bag of n-grams and depicted as an occurrence frequency vector, ignoring information contained in n-grams of original text. As a result, vector has an excessive numeral of unnecessary and duplicated characteristics. In this research, they suggest a novel approach to use N-

gram models for feature extraction from phrase level analysis of sentiment. One of the three-word N-grams (N = 3) that comprise a phrase is chosen to serve as an emotion term. Consequently, after producing the three word N-grams, they activate a sentiment lexicon that directs the user to an N-gram that has the emotion term in touch.

The detected N-gram is divided into three words. From the three words, their Part of Speech POS tags, and their sentiment orientations, 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.

To get feeling aspects (A), combine words, POS tags, and semantic orientations as follow;

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

3.2.2. Audio features are extracted with Mel-Frequency Cepstral Coefficients (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.

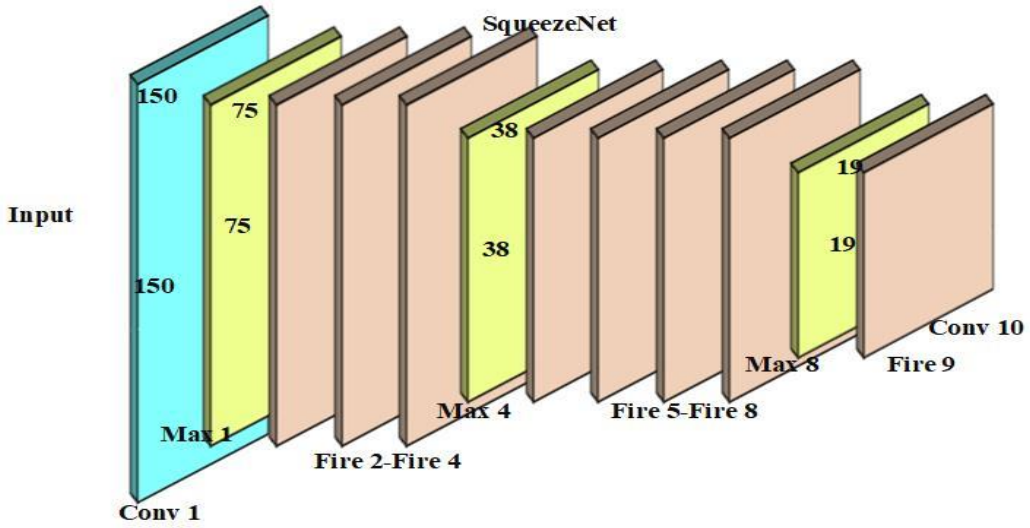


Figure 3: Architecture of Squeezenet.

- 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.
- $r : 2^\Theta \to [0,1]$ 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:

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

Where,

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

Combine Result with Text:

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

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

3.4. Capsule Net

One capsule can hold several other capsules in a network of capsules. The classification network uses the capsule module to represent the capsule orientation and classification probability to describe the different orientations of the emotions. Therefore, Capsule networks are more enriched and efficient than traditional neural network models, including CNN's. A capsule network produces emotions rather than a scalar value like a CNN pooling layer. The above fused data x is the input of capsule net [25].

This is accomplished by passing the final data x_m to the fused data, which represents the output of a fused data.

$$\hat{e}_{j|i} = w_t^{cap} e_i \quad (21)$$

A non-linear activation function is used to convert the final state x_m of fused data into a feature capsule e_i. A few more specifics are in the code, e_i determines the correlation between the input and output layers, and $\hat{e}_{j|i}$ is used to predict outputs, where w_t^{cap} is the weight for the input.

$$g_j = \sum_{i=1}^{n} a_{ij} \hat{e}_{j|i} \quad (22)$$

The coupling coefficients a_{ij} are calculated using the dynamic routing process. This process ignores input that contains trivial and irrelevant elements. Audio, video, and text are weighted according to the coupling coefficient a_{ij} .The weight of a feature is higher when it has a high a_{ij} value and vice versa. Capsule outputs g_j are calculated by summation of all prediction feature maps.

$$g_j = \sum_{i=1}^{n} a_{ij} \hat{e}_{j|i} \quad (23)$$

By using the below equation, the softmax function calculates the coupling coefficient a_{ij} .

$$a_{ij} = \frac{\exp(z_{ij})}{\sum_e \exp(z_{ie})} \quad (24)$$

The following equation updates z_{ij} .This layer represents the capsule network of higher layers.

$$z_{ij} \leftarrow z_{ij} + R_j^T f(e_i, \theta_j) \quad (25)$$

With this equation, the final output is r_j normalized using a squash function (an activation function of non-linear) that accounts for different orientations.

$$r_j = \frac{\|g_j\|^2}{1+\|g_j\|^2} \frac{g_j}{\|g_j\|} \quad (26)$$

Consequently, the size of internal product of r_j and $\hat{v}_{j|i}$ determines (agrees) which capsule in the following layer will probably route.

$$y_{ij} = r_j . \hat{v}_{j|i} \quad (27)$$

The profit for Capsule Networks is the total of all category capsule losses, each of which is computed as a distinct margin loss P_k , for each category capsule k.

$$P_k = H_k \max(0, m^+ - \|r_k\|)^2 + \lambda(1-H_k)\max(0, \|r_k\| - m^-)^2 \quad (28)$$

where H_k indicates that the instantiation is present in category capsule k; and m^+, m^- and λ hyper-parameters that regulate the existence-based loss value.

3.4.1. Hyper-parameter optimization using SCSO algorithm

Hyper-parameter optimization using the SCSO algorithm involves leveraging the collective behavior of sand cats to explore and exploit the hyper-parameter space efficiently. The SCSO algorithm iteratively adjusts the hyper-parameters by mimicking the hunting and foraging patterns of sand cats, balancing exploration and exploitation. Each candidate solution (sand cat) updates its position based on the best-found solutions and random perturbations to avoid local optima. The algorithm continues until a stopping criterion, such as a maximum number of iterations or convergence threshold, is met. SCSO has shown promise in effectively tuning hyper-parameters for various machine learning models.

Table 1: range of hyper parameter initialization

Hyper-parameters	**Range**
Number of Capsules (NoC)	[8, 16, 32]
Dimensions of Capsules (DoC)	[8, 16, 32]
Learning rate (LR)	[0.001, 0.01, 0.1]

Batch size (BS)	[32, 64, 128]
Dropout rate (DR)	[0.1, 0.25, 0.5]
Reconstruction loss weight (RLW)	[0.0005, 0.005, 0.05]

The SCSO is used optimize these hyper parameters. The SCSO [26] algorithm's step-by-step procedure is explained below.

Step 1: Initialization: The primary goal of this approach is to select the optimal Hyper-parameter. First, ascertain the amount of repetitions, the dimensionality of the variables D, and the size of the Sand Cat Swarm N. In this case, the Sand Cat Swarm represents the hyper-parameter quantity. Table 1 lists the Hyper-parameter utilized in the capsule net. At first, the Sand Cat Swarm is selected at random. Solutions are first filled in as the following:

$$P_N = \{X_1, X_2, ..., X_N\} \quad (29)$$

Here, X_N is the Nth solution or Sand Cat Swarm's location

$$X_i = \{\text{NoC, DoC, LR, BS, DR, RLW}\}_i \quad (30)$$

Step 2: Opposite solution generation: opposing solutions are generated once the solution has been initialized. This stage is employed to optimize the algorithm's search capability. Whereas $X_i \in [a,b]$ be a genuine figure, the opposite answer X_i' can be calculated as follows;

$$X_i' = a + b - X_i \quad (31)$$

Step 3: Fitness calculation: Following the initial setup phase, each solution's fitness is assessed by the suggested SCSO approach. The minimal fitness value is considered to be the optimum answer. Equation (14) is utilized in the computation of fitness.

$$Fitness = Max\left(\frac{TP+TN}{TP+TN+FP+FN}\right) \quad (32)$$

Step 4: Updating with SCSO: SCSO utilizes 3 distinct techniques known as searching the prey (exploration), Attacking on the prey (exploitation) and Exploration and exploitation.

Strategy 1: Searching the prey (exploration)

The SCSO method's search method is explained. Sand cats use minimal noise production as a prey seeking method. For every sand cat, the answer is expressed as $D_i = (D_{i1}, D_{i2}, D_{i3}, ..., D_{ix})$. The capacity to hear helps the SCSO algorithm discover sand cats at frequencies that are low.

$$\vec{j}_s = G_M - \left(\frac{2 \times G_M \times iter_c}{iter_{\max} + iter_{\max}}\right) \quad (33)$$

$$\vec{J} = 2 \times \overrightarrow{j_S} \times rand(0,1) - \overrightarrow{j_S} \quad (34)$$

Therefore, the $\overrightarrow{j_S}$ shows the whole sensitivity range, which drops linearly from 2 to 0. Besides, $\vec{j}$ shows the range of sensitivity for each cat. The $\vec{j}$ is employed for tasks throughout the periods of exploration or exploitation, $\vec{j}_s$, controls the transitions in these phases by guiding the J parameter. In addition, $iter_c$ is present iteration and $iter_{\max}$ is maximum iterations.

$$\vec{j} = \overrightarrow{j_S} \times rand(0,1) \quad (35)$$

According to the position of the top candidate, every search agent (sand cat) updates the position it holds $\overrightarrow{pos_{bc}}$ and its present position $\overrightarrow{pos_{cc}}$ and its understanding range $(\vec{y})$. This causes algorithm must have a minimal cost of operating and a low complexity.

$$\overrightarrow{Pos}(h+1) = \vec{y}.\left(\overrightarrow{Pos_{bc}}(h) - rand(0,1).\overrightarrow{Pos_c}(h)\right) \quad (36)$$

Strategy 2: Attacking on the prey (exploitation)

As previously stated, the sand cats use their hearing to identify their prey. In order to computationally simulate the SCSO assaulting phase, there is a gap between the optimal position ($\overrightarrow{Pos_b}$) (best solution) and the current position ($\overrightarrow{Pos_v}$) of the sand cat.

$$\overrightarrow{Pos_{rnd}} = \left|rand(0,1).\overrightarrow{Pos_b}(h) - \overrightarrow{Pos_c}(h)\right|,$$
$$\overrightarrow{Pos}(h+1) = \overrightarrow{Pos_b}(h) - \vec{j}.\overrightarrow{Pos_{rnd}}.\cos(\theta) \quad (37)$$

Strategy 3: Exploration and exploitation

The adaptable principles that guide exploration and exploitation include j_s and J parameters. SCSO can smoothly transition between two stages thanks to these factors. Given that the J parameter is dependent upon j_s, its fluctuation range will be also decreased.

$$\vec{D}(h+1) = \begin{cases} \overrightarrow{Pos_b}(h) - \overrightarrow{Pos_{rnd}}.\cos(\theta).\vec{j} & |J \le 1; \exp laitation| \\ \vec{j}.\left(\overrightarrow{Pos_{bc}}(h) - rand(0,1).\overrightarrow{Pos_c}(h)\right) & |J| > 1; \exp loration \end{cases} \quad (38)$$

Step 5: Termination condition: Until the optimal hyper-parameter is selected, the process continues. The capsule net receives the chosen hyper parameter value.

Algorithm 1: Sand Cat Swarm Optimization Algorithm Pseudocode

Set up the population initially.

Determine the fitness function by utilizing the objective function as a basis.

Initialize the parameters j, j_s and J.

While h ≤ maximum iteration:

For every search engine

Determine the angle θ at random by using the roulette wheel choice ($0^o \leq \theta \leq 360^o$)

If $|J| \leq 1$

Adjust the search agent's location in light of Equation (19)

Else

Adjust the search agent's location in light of Equation (18)

end

h=h++

end

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.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (39)$$

$$recall = \frac{TP}{TP + FN} \quad (40)$$

$$precision = \frac{TP}{TP + FP} \quad (41)$$

$$F1-score=\frac{2TP}{2TP+FP+FN} \quad (42)$$

$$FPR=\frac{F_P}{T_N+F_P} \quad (43)$$

$$FNR=\frac{F_N}{T_P+F_N} \quad (44)$$

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

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

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 1400 exchanges and 13,000 words from the Friends series on television. Several speakers took part in the conversations. Any one among the following seven emotions—Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear—has been assigned to each statement made in a discussion. Each speech in MELD additionally includes a sentiment descriptor (good, negative, or neutral) [27].

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, CNN, and SVM. 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.

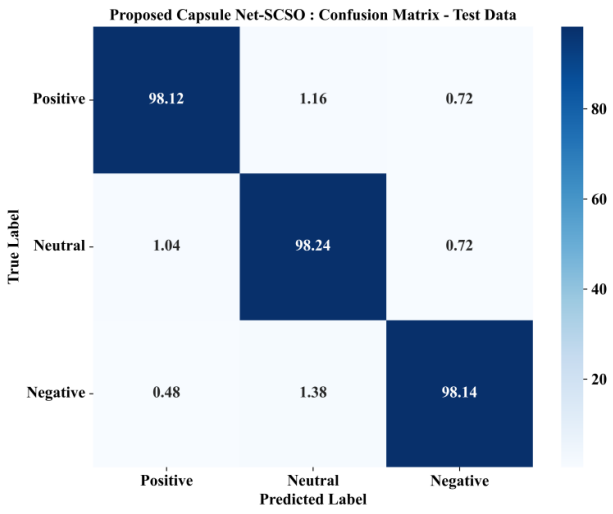
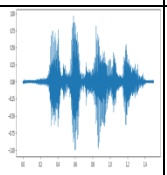
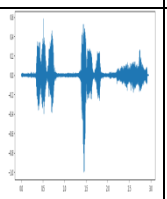
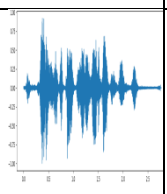


Figure 4: Confusion Matrix

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						

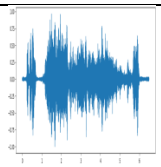
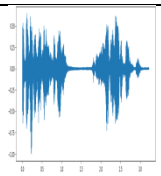
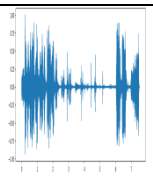
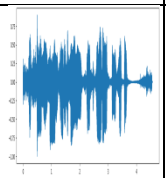
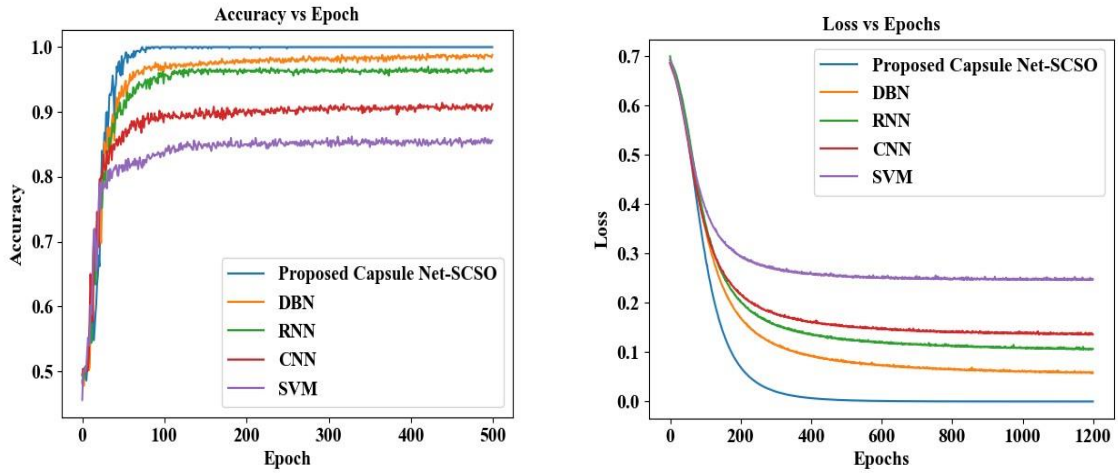
	startled me.						
Joy	Compared to our first outing, this is so much better.						
Sadness	I had to leave. It's getting really late now, but I love and miss you too.						
Neutral	Hey, what about the kangaroo scene? Did you enjoy that section?						
Disgust	Both my granny and her new partner are a little 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: Accuracy Vs Loss Curve Comparison

Figure 6 compares the concert of the proposed Capsule Net-SCSO model with DBN, RNN, CNN, and SVM models in terms of accuracy and loss over epochs. The accuracy curve shows that the proposed Capsule Net-SCSO achieves the highest accuracy faster than other models, stabilizing around 100 epochs. The loss curve indicates that the Capsule Net-SCSO model converges more quickly and to a lower loss value compared to the other models, demonstrating its superior learning efficiency and effectiveness in minimizing error over training epochs.

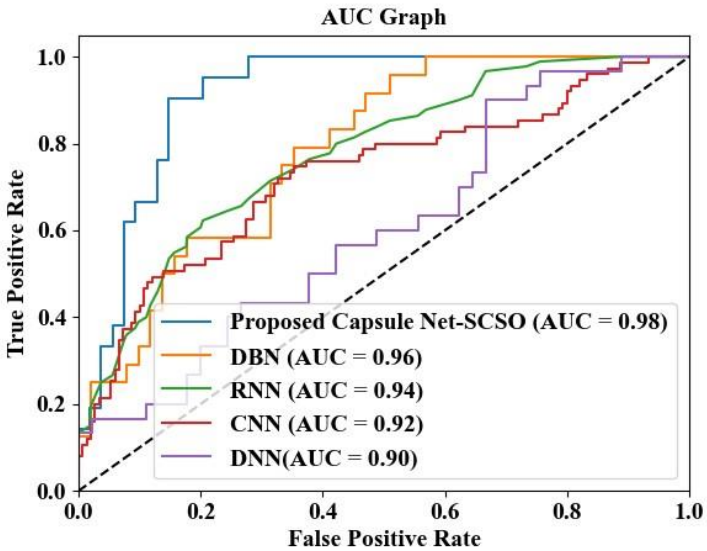


Figure 7: AUC Graph

The above figure 7 shows AUC graph compares show of future Capsule Net-SCSO model with DBN, RNN, CNN, and DNN models. The proposed Capsule Net-SCSO achieves the highest AUC of 0.98, indicating superior performance in distinguishing between classes. DBN follows with an AUC of 0.96, while RNN, CNN, and DNN achieve AUCs of 0.94, 0.92, and 0.90, respectively. The results demonstrate that the Capsule Net-SCSO model has the best true positive rate against the false positive rate, highlighting its effectiveness and reliability in classification tasks compared to other models.

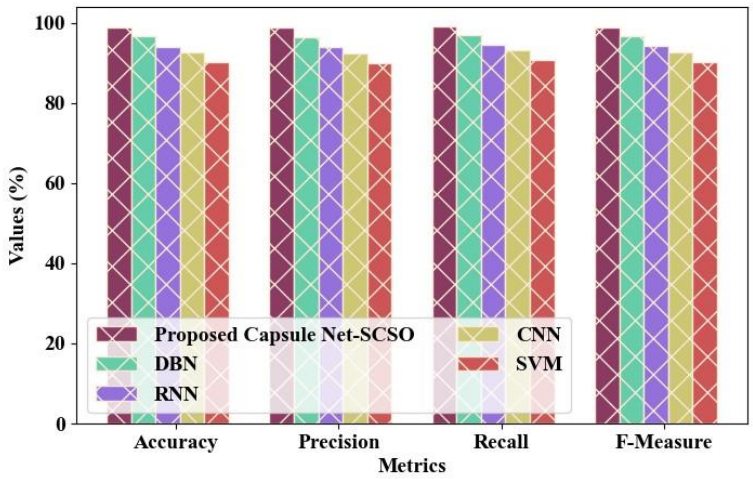


Figure 8: classification comparison with existing method

The figure 8 classification comparison graph illustrates the performance of the proposed Capsule Net-SCSO model against DBN, RNN, CNN, and SVM across four metrics: accuracy, precision, recall, and F-measure. The Capsule Net-SCSO 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 Capsule Net-SCSO model compared to traditional methods in classification tasks.

Table 2: Comparison table

Methods	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Proposed Capsule Net-SCSO	98.91	98.83	99.04	98.44
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
SVM	90.1	89.85	90.75	90.3

Table 2 shows the comparative performance of various methods, highlighting the superiority of the proposed Capsule Net-SCSO model. It achieves the highest accuracy (98.91%), precision (98.83%), recall (99.04%), and F-measure (98.44%), outperforming DBN, RNN, CNN, and SVM. DBN is the next best, with accuracy and F-measure scores of 96.62% and 96.66%, respectively. RNN, CNN, and SVM follow with progressively lower scores, demonstrating that the Capsule Net-SCSO model excels in all evaluated metrics, making it the most effective and reliable method for the classification tasks in this study.

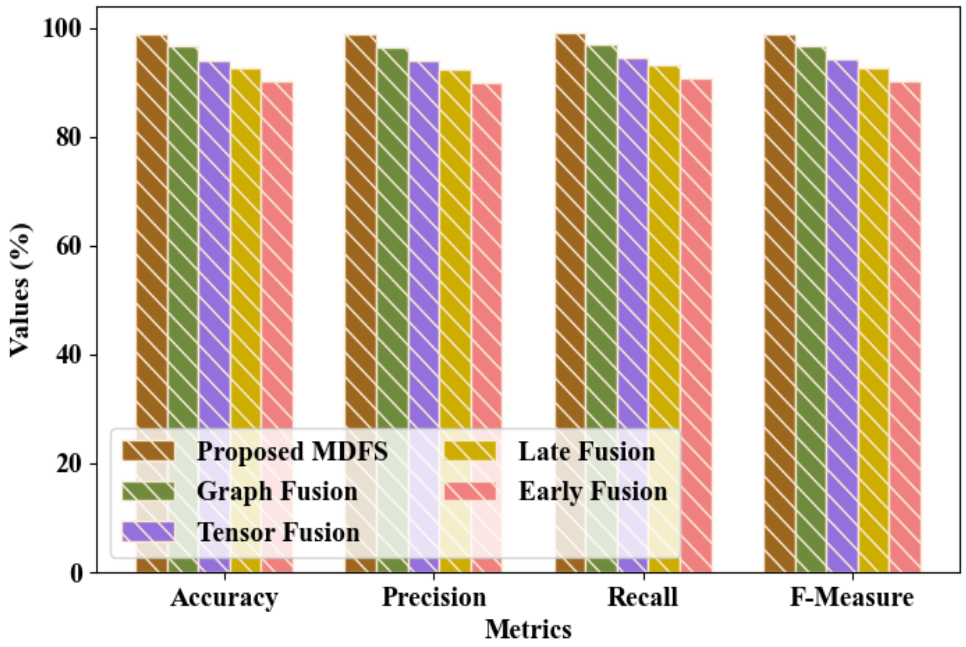


Figure 9: Feature fusion comparison

Figure 9 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 3: 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 3 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.

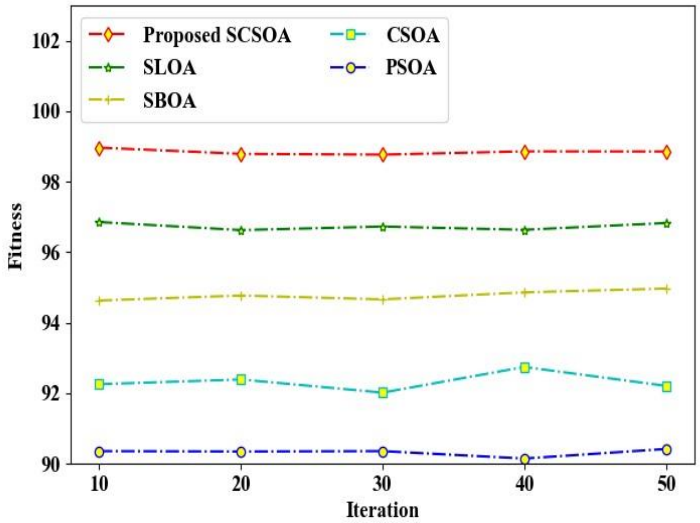


Figure 10: Fitness Improvement over Iteration Graph

In the result section, Figure 10 depicts the fitness improvement over iterations, showing a clear upward trend indicative of the optimization process's effectiveness. The graph reveals a steady increase in fitness scores as iterations progress, suggesting iterative refinement and convergence towards better-performing solutions. This visual representation underscores the algorithm's efficacy in enhancing performance metrics, validating the approach's success in achieving improved outcomes through successive iterations.

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 4: Performance Comparison 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 electroencephalogram (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	Capsule Net	98.91	98.83	99.04	98.94

Table 4 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

This research effectively demonstrates the potential of a comprehensive multimodal approach for emotion recognition by integrating audio, video, and text data. The proposed feature extraction methods, including MFCC for audio, Squeezenet for visual data, and the Assimilated N-gram Approach for text, significantly enhance the quality of the input features. By employing a robust

multimodal data fusion strategy and utilizing Capsule Net for classification, our approach achieved impressive performance metrics, with an accuracy of 98.91%, precision of 98.83%, recall of 99.04%, and F-measure of 98.94. These results underscore the effectiveness of our model in accurately identifying emotional cues, making it a valuable contribution to the field of multimodal emotion recognition. Future research could explore the integration of additional modalities, such as physiological signals, to further enhance emotion recognition accuracy. Additionally, developing real-time processing capabilities for deployment in applications like mental health monitoring and interactive systems would be beneficial.

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," *Mathematical Problems in Engineering*, 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," *Computational Intelligence and Neuroscience*, vol. 2023, pp. 1–18, Jan. 2023, doi: 10.1155/2023/9645611.

[25] F. Kinli and F. Kiraç, "FashionCapsNet: Clothing Classification with Capsule Networks," *Bilişim Teknolojileri Dergisi*, vol. 13, no. 1, pp. 87–96, Jan. 2020, doi: 10.17671/gazibtd.580222.

[26] A. Seyyedabbasi and F. Kiani, "Sand Cat swarm optimization: a nature-inspired algorithm to solve global optimization problems," *Engineering with Computers*, vol. 39, no. 4, pp. 2627–2651, Aug. 2023, doi: 10.1007/s00366-022-01604-x.

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

Author's Profile:

Sanjeeva Rao Sanku[1]: is pursuing Ph.D in Computer Science and Engineering from Osmania University, Hyderabad. He completed his M.Tech and B.Tech in CSE from JNTU Hyderabad University. His research areas include Data mining, Machine Learning and Deep Learning. He has 17+ years of teaching and research experience.

B.Sandhya[2]: was awarded Ph.D in Computer Science from University of Hyderabad in the year 2011. She is currently working as Professor in CSE department in MVSR Engineering college, Hyderabad. She has 22+ years of teaching, research and consultancey experience. Her principal areas of research include Image Processing, Machine learning, Deep learning and Computer Vision. She has authored/co-authored around 40 research publications in international conferences and journals, with total citations of about 170.