OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

Emotion Recognition using Attention Mechanism: An Empirical Study Across Diverse Data Modalities

Nikita Joshi[1],Dr. Rakesh Kumar Khare[2]
M. Tech. Scholar,Dept. of CSE,SSIPMT, Raipur [1]
Associate Professor,Dept. of CSE,SSIPMT, Raipur[2]

Abstract: ***This research investigates the effectiveness of attention mechanisms for emotion recognition across three distinct modalities: textual expressions, facial imagery, and vocal patterns. We implemented separate attention-enhanced models for each modality to systematically compare their performance and identify which types of emotional data benefit most from attention-based processing. Our methodology employed three benchmark datasets: the Emotions dataset for NLP for text analysis, FER2013 for facial expression recognition, and RAVDESS for speech emotion recognition. Each modality utilized tailored preprocessing pipelines and attention mechanisms designed to focus on emotionally relevant features within their respective domains. Experimental results revealed differential effectiveness across modalities, with speech emotion recognition achieving the highest accuracy of 38.1%, followed by text-based recognition at 32.1%, and facial expression recognition at 29.8%. While these results demonstrate the feasibility of applying attention mechanisms to emotion recognition tasks, they also highlight significant performance gaps compared to established benchmarks in the field. The findings suggest that attention mechanisms show promise for emotion recognition, particularly in speech analysis, but require substantial architectural improvements and enhanced feature extraction techniques to achieve competitive performance levels for practical deployment..***

Keywords: Emotion Recognition, Attention Mechanism, Multimodal Analysis, Deep Learning, Affective Computing

---

I. INTRODUCTION

1.1 Background and Context

The ability to recognize and interpret human emotions represents one of the most fundamental aspects of human communication and social interaction. As artificial intelligence systems become increasingly integrated into daily life, the development of robust emotion recognition capabilities has emerged as a critical research frontier with profound implications for human-computer interaction, mental health monitoring, educational technology, and social robotics. The complexity of human emotional expression, which manifests through multiple channels including facial expressions, vocal patterns, and textual communication, presents both opportunities and challenges for computational approaches to emotion understanding.

Traditional emotion recognition systems have typically focused on individual modalities, developing specialized algorithms for processing either textual sentiment, facial expressions, or speech patterns. While these approaches have achieved considerable success within their respective domains, they often fail to capture the nuanced and multifaceted nature of human emotional expression. The emergence of attention mechanisms in deep learning has opened new possibilities for emotion recognition by enabling systems to automatically identify and focus on the most emotionally relevant features within complex data streams.

Attention mechanisms, originally developed for machine translation and natural language processing tasks, have demonstrated remarkable success in learning to selectively focus on relevant information while ignoring irrelevant distractions. This capability aligns naturally with human emotional processing, where individuals instinctively attend to specific words, facial features, or vocal characteristics that convey emotional meaning. The application of attention mechanisms to emotion recognition represents a promising direction for developing more sophisticated and human-like emotional understanding systems.

1.2 Research Motivation

The motivation for this research stems from the recognition that while attention mechanisms have shown tremendous success in various artificial intelligence applications, their specific effectiveness for emotion recognition across different modalities remains largely unexplored. Current emotion recognition systems often struggle with the inherent variability and context-dependency of emotional expression, leading to inconsistent performance across different individuals, situations, and expression styles. The ability of attention mechanisms to dynamically focus on relevant features suggests they could address many of these limitations.

Furthermore, the lack of systematic comparative studies examining attention mechanism performance across different

emotional modalities represents a significant gap in the current literature. While individual studies have applied attention to specific emotion recognition tasks, there is limited understanding of how these mechanisms perform comparatively across textual, visual, and auditory emotional data. This knowledge gap hinders the development of informed strategies for emotion recognition system design and deployment.

The increasing demand for emotion-aware applications in healthcare, education, marketing, and entertainment sectors further motivates this research. Healthcare applications require accurate emotion monitoring for mental health assessment and therapeutic intervention. Educational technologies benefit from understanding student emotional states to adapt learning experiences. Marketing applications seek to gauge consumer emotional responses to products and advertisements. Entertainment systems aim to create more engaging and emotionally responsive experiences. All of these applications would benefit from a deeper understanding of how attention mechanisms can enhance emotion recognition capabilities.

1.3 Research Problem Statement

Despite the theoretical promise of attention mechanisms for emotion recognition, several fundamental questions remain unanswered. First, it is unclear which types of emotional data benefit most from attention-based processing. Different modalities present distinct characteristics that may interact differently with attention mechanisms. Textual emotional expression relies heavily on semantic relationships and contextual meaning, facial expressions depend on spatial feature patterns and subtle visual cues, while vocal emotional expression involves temporal sequences and spectral characteristics. The effectiveness of attention mechanisms may vary significantly across these different data types.

Second, the optimal architectural design for attention-based emotion recognition systems remains poorly understood. While basic attention mechanisms provide a starting point, the specific configurations, parameter settings, and integration strategies that maximize emotion recognition performance are not well established. The trade-offs between computational efficiency and recognition accuracy in attention-based systems also require careful investigation.

Third, the performance expectations for attention-based emotion recognition systems lack clear benchmarks. Without systematic evaluation against established standards, it is difficult to assess whether attention mechanisms represent a genuine advancement in emotion recognition or merely an alternative approach with similar limitations to existing methods. This uncertainty impedes the development of practical applications and the allocation of research resources.

II.LITERATURE REVIEW

Emotion recognition using text, speech, and face image data has seen significant advancements, driven by traditional machine learning and deep learning techniques. While each modality has its strengths, challenges such as data limitations, emotional ambiguity, and cross-domain adaptability remain. Future research should focus on developing multimodal approaches that combine these data types to enhance robustness and accuracy in real-world applications.

2.1 Text-Based Emotion Recognition

Text-based emotion recognition involves analyzing written content to identify emotional states. Traditional machine learning approaches, such as multinomial Naive Bayes (MNB), Support Vector Machines (SVM), and Random Forests, have been widely used for text-based emotion recognition. These methods rely on handcrafted features like TF-IDF and keyword-based approaches (Abdykerimova et al., 2024) (Shah et al., 2024) (Kashif et al., 2016).

Deep learning techniques, such as Long Short-Term Memory (LSTM) networks and Transformers, have shown superior performance in text-based emotion recognition. These models leverage word embeddings like Word2Vec and GloVe to capture semantic and syntactic information. For instance, bidirectional LSTMs (BiLSTM) have been effective in capturing contextual relationships in text, achieving high accuracy in emotion classification tasks (Chaithra & Samanvaya, 2024) (Asghar et al., 2022) (Deng & Ren, 2021).

Recent advancements include the use of semantic-emotion neural networks (SENN), which combine semantic and emotional information through sub-networks. These models have demonstrated improved performance over traditional approaches, particularly in capturing nuanced emotional expressions (Batbaatar et al., 2019).

Challenges in Text-Based Emotion Recognition

- **Ambiguity and Context Dependency**: Textual data often contains implicit emotions, sarcasm, and metaphorical language, making it challenging to accurately detect emotions (Seyeditabari et al., n.d.).
- **Limited Training Data**: The availability of large-scale, high-quality datasets for specific emotions remains a challenge (Deng & Ren, 2021).
- **Cross-Domain Adaptability**: Models trained on one dataset often struggle to generalize to other domains or languages (Shah et al., 2024).

2.2 Speech-Based Emotion Recognition

Speech-based emotion recognition focuses on extracting emotional cues from audio signals. Traditional approaches involve acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs), and spectral features like zero-crossing rate (ZCR) and pitch. These features are often combined with classifiers like Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) ("Enhancing Speech Emotion Recognition Through Advanced Feature Extraction and Deep Learning: A Study Using the RAVDESS Dataset", 2024) (Bhangale & Kothandaraman, 2023) ("Speech Emotion Classification: A Survey of the State-of-the-Art", 2023).

Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have revolutionized speech emotion recognition. CNNs are effective in extracting spectral and temporal features from speech signals, while RNNs capture long-term dependencies in emotional expressions. For example, 1-D CNNs have been used to minimize computational complexity while maintaining high

accuracy (Bhangale & Kothandaraman, 2023) ("Speech Emotion Classification: A Survey of the State-of-the-Art", 2023).

Challenges in Speech-Based Emotion Recognition

- **Noise Robustness**: Background noise and variations in recording conditions can degrade performance (Bhangale & Kothandaraman, 2023).
- **Emotional Ambiguity**: Overlapping acoustic features between similar emotions (e.g., anger and frustration) make classification challenging ("Speech Emotion Classification: A Survey of the State-of-the-Art", 2023).
- **Cultural and Linguistic Variability**: Emotional expression varies across cultures and languages, requiring diverse datasets for robust models (Drakopoulos et al., 2019).

2.3 Face Image-Based Emotion Recognition

Facial emotion recognition involves analyzing facial expressions to identify emotional states. Traditional methods rely on handcrafted features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and facial landmarks. These features are often classified using SVMs and Random Forests (Jain & Kavita, 2024) (Gupta et al., 2024).

Deep learning, particularly Convolutional Neural Networks (CNNs), has significantly advanced facial emotion recognition. CNNs automatically extract spatial features from facial images, achieving high accuracy in classifying emotions. Models like ResNet50 and InceptionV3 have been widely adopted for their effectiveness in capturing intricate facial features (Jain & Kavita, 2024) (Gupta et al., 2024) (Mahastama et al., 2024).

Challenges in Face Image-Based Emotion Recognition

- **Pose and Illumination Variations**: Changes in head pose, lighting conditions, and occlusions (e.g., glasses or masks) can affect accuracy (Jain & Kavita, 2024) (Gupta et al., 2024).
- **Emotional Intensity**: Recognizing subtle emotional expressions remains a challenge (Khan, 2022).
- **Dataset Limitations**: The need for large, diverse datasets with balanced emotion categories is a persistent issue (Singh & Shukla, 2023).

Table 1 Comparative Analysis of Emotion Recognition Modalities

Modality	Key Methods	Challenges	Datasets
Text	Traditional: TF-IDF, SVM, Random Forest	Ambiguity, sarcasm, limited datasets	IEMOCAP
	Deep Learning: LSTM, BiLSTM, Transformers		Ekman's six emotions
Speech	Traditional: MFCC, LPCC, HMM, GMM	Noise, emotional ambiguity, cultural variability	RAVDESS
	Deep Learning: CNN, RNN		EMODB
Face	Traditional: LBP, HOG, SVM	Pose, illumination, subtle expressions	
	Deep Learning: CNN, ResNet50, InceptionV3		CK+, IMED

III.DATASET

This research employs three carefully selected benchmark datasets representing different modalities for emotion recognition tasks. The datasets encompass textual expressions, facial manifestations, and vocal emotional characteristics, providing a comprehensive foundation for developing and evaluating the proposed attention-based emotion recognition framework. Each dataset has been chosen for its established credibility in the research community and its alignment with the seven fundamental emotion categories that form the basis of this study.

3.1 Text Emotion Dataset

The textual emotion dataset utilized in this research is sourced from Kaggle's "Emotions dataset for NLP" collection, which serves as a foundational resource for natural language processing classification tasks. This dataset contains thousands of English text samples that have been manually annotated with emotional labels, making it particularly suitable for training and evaluating text-based emotion recognition systems.

The dataset encompasses a diverse range of textual expressions, including social media posts, conversational snippets, and written statements that naturally exhibit emotional content. Each text sample has been carefully labeled by human annotators to ensure accuracy and consistency in emotion classification. The textual data represents authentic human expressions across various contexts, from casual conversations to more formal written communications, providing the model with exposure to different linguistic patterns and emotional expressions.

```
im feeling rather rotten so im not very ambitious right now;sadness
im updating my blog because i feel shitty;sadness
i never make her separate from me because i don t ever want her to feel like i m ashamed with her;sadness
i left with my bouquet of red and yellow tulips under my arm feeling slightly more optimistic than when i ar
i was feeling a little vain when i did this one;sadness
i cant walk into a shop anywhere where i do not feel uncomfortable;fear
i felt anger when at the end of a telephone call;anger
i explain why i clung to a relationship with a boy who was in many ways immature and uncommitted despite the
i like to have the same breathless feeling as a reader eager to see what will happen next;joy
i jest i feel grumpy tired and pre menstrual which i probably am but then again its only been a week and im
i don t feel particularly agitated;fear
i feel beautifully emotional knowing that these women of whom i knew just a handful were holding me and my t
i pay attention it deepens into a feeling of being invaded and helpless;fear
i just feel extremely comfortable with the group of people that i dont even need to hide myself;joy
i find myself in the odd position of feeling supportive of;love
i was feeling as heartbroken as im sure katniss was;sadness
i feel a little mellow today;joy
```

Figure 1 Snippet from Emotions dataset for NLP

What makes this dataset particularly valuable for research purposes is its balanced representation across multiple emotion categories. The text samples vary in length from short phrases to longer sentences, allowing the attention mechanism to learn from both concise emotional expressions and more elaborate emotional narratives. The dataset includes contemporary language patterns and colloquialisms that reflect modern communication styles, ensuring the model's relevance to current real-world applications.

The preprocessing pipeline for this dataset involves standard natural language processing techniques, including tokenization, normalization, and encoding procedures that prepare the text for embedding generation. The dataset's structure allows for straightforward integration with various text embedding models, making it adaptable to different feature extraction approaches within the attention-based framework.

3.2 Facial Expression Dataset (FER2013)

The facial expression component of this research utilizes the FER2013 dataset, a widely recognized benchmark in the computer vision community for learning facial expressions from images. This dataset originated from a Kaggle challenge focused on facial expression recognition and has since become a standard evaluation benchmark for emotion recognition systems.

The FER2013 dataset contains grayscale facial images classified into seven distinct emotional categories: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Each image in the dataset is standardized to 48x48 pixels, providing consistent input dimensions for neural network processing while maintaining sufficient resolution to capture essential facial features and micro-expressions that are crucial for emotion recognition.

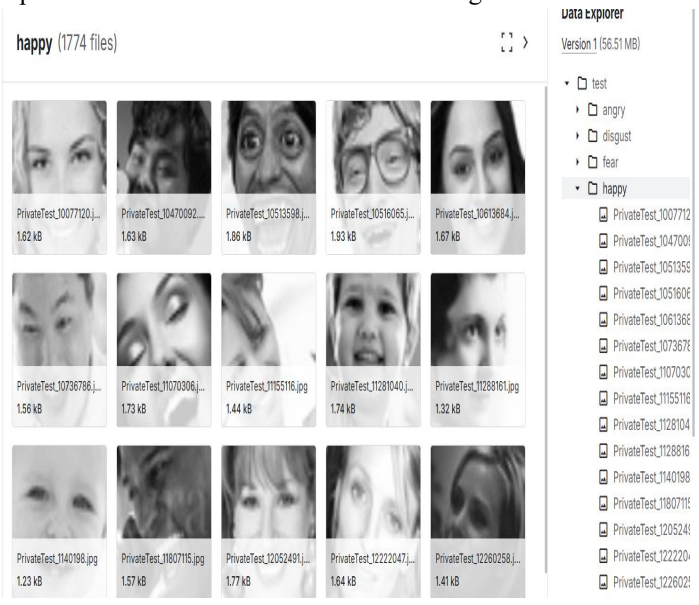

Figure 2 Snippet of FER2013

The dataset encompasses facial expressions captured under various lighting conditions, angles, and demographic representations, contributing to its robustness for real-world applications. The images include both posed and spontaneous expressions, offering a comprehensive view of how emotions manifest in human facial features. This diversity ensures that the attention mechanism can learn to focus on relevant facial regions regardless of variations in image quality, lighting, or facial orientation.

One of the strengths of the FER2013 dataset lies in its extensive validation by the research community. Multiple studies have demonstrated the effectiveness of various approaches on this dataset. The dataset's challenging nature, stemming from its real-world variability and the inherent difficulty of emotion recognition from static images, makes it an excellent testbed for evaluating the proposed attention mechanism's ability to identify and weight relevant facial features.

The preprocessing pipeline for FER2013 includes standard image processing techniques such as normalization, contrast enhancement, and data augmentation procedures that increase the dataset's effective size while maintaining the integrity of emotional expressions. The standardized format facilitates seamless integration with convolutional neural networks and other image processing architectures within the attention-based framework.

3.3 Speech Emotion Dataset (RAVDESS)

The audio component of this research leverages the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which serves as a comprehensive emotional speech dataset for machine learning applications. RAVDESS represents a dynamic, multimodal collection containing 7,356 validated presentations of emotion, making it one of the most extensive and rigorously validated emotional speech databases available.

The dataset contains 3-second audio clips of the same two sentences spoken by 24 different actors, covering an emotional range of 7 emotions, with 12 male and 12 female actors providing a total of 1,440 samples. This balanced gender representation ensures that the emotion recognition system can effectively learn from diverse vocal characteristics and speaking patterns, reducing potential bias in the model's performance across different demographic groups.

The RAVDESS dataset employs a systematic naming convention where each audio file encodes specific information about the vocal channel (speech or song), emotion categories (neutral, calm, happy, sad, angry, fearful, disgust, surprised), and emotional intensity levels (normal or strong intensity). This detailed categorization allows for fine-grained analysis of emotional expressions and provides researchers with the flexibility to focus on specific emotional intensities or vocal characteristics.

The dataset's structure includes numbered identifiers for different aspects:

odd-numbered actors represent male speakers while even-numbered actors represent female speakers, providing clear demographic information that can be utilized for analysis. The professional quality of the recordings, performed by trained actors under controlled conditions, ensures consistent audio quality and authentic emotional expression across all samples.

The RAVDESS dataset's applications extend beyond academic research into practical domains such as medical field applications and customer call center systems, demonstrating its real-world relevance. The dataset's validation process and the involvement of professional actors in creating emotionally authentic expressions contribute to its reliability as a benchmark for speech emotion recognition systems.

For this research, the RAVDESS dataset undergoes preprocessing that includes audio normalization, feature extraction procedures for generating spectrograms and mel-frequency cepstral coefficients, and segmentation techniques that prepare the audio data for integration with the attention mechanism. The consistent duration of audio clips facilitates batch processing and ensures uniform input dimensions for the neural network architecture.

IV.METHODOLOGY

Our research introduces an attention-based approach to emotion recognition applied across three distinct modalities: text, facial expressions, and voice.

We developed separate attention mechanisms for each modality to understand what matters most in each type of emotional expression. This approach mirrors how humans naturally process emotions by paying attention to the most relevant cues, whether that's a particular word in a sentence, a specific facial feature, or a tone of voice.

Our framework applies consistent attention principles across three different emotion recognition tasks. Instead of building traditional models for text, face, and speech analysis, we created attention-enhanced systems that can identify the most emotionally relevant features within each modality. This allows each system to learn what patterns matter most for emotion detection in its respective domain.

4.1 Transforming Text into Meaningful Representations

Working with the Emotions Dataset for NLP presented us with the challenge of converting human language into something a computer can understand while preserving the emotional nuances that make text meaningful. We started with raw text samples that varied greatly in style, length, and formatting - much like the diverse ways people express themselves in real life.

Our first step involved cleaning and standardizing the text data. We approached this carefully, recognizing that some elements that might seem like "noise" actually carry emotional weight. For instance, exclamation marks and capitalization often indicate strong emotions, so we preserved these while removing truly irrelevant elements like URLs and special characters that don't contribute to emotional understanding.

The tokenization process broke down each text sample into individual meaningful units. A phrase like "I'm absolutely thrilled about this opportunity!" becomes a sequence of tokens that our system can process individually while maintaining their relationship to one another. This transformation is crucial because it allows our attention mechanism to focus on specific words that carry the most emotional significance.

After tokenization, we had sequences of varying lengths - some short emotional outbursts, others longer, more complex expressions.

We used BERT embeddings to convert these tokens into rich numerical representations that capture not just the meaning of individual words but also their emotional context within the sentence. This resulted in each text sample being represented as a matrix where each row contains a 768-dimensional vector capturing the contextual meaning of a single word.

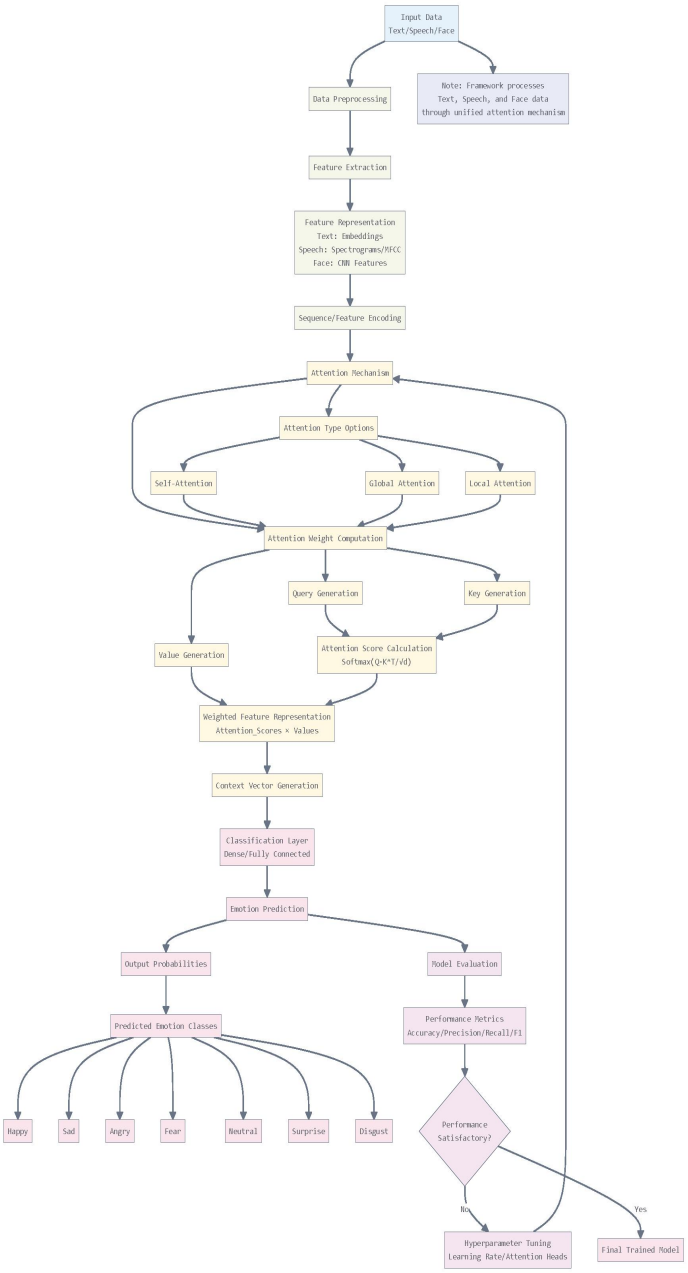

Figure 3 Methodology Flowchart

4.2 Processing Facial Expressions from Images

The FER2013 dataset brought its own unique challenges, as facial expressions are incredibly nuanced and can vary dramatically based on lighting, angle, and individual differences in how people show emotions. We received 48x48 pixel grayscale images that needed careful preprocessing to ensure our system could learn meaningful patterns rather than being confused by variations in image quality or lighting conditions.

Our preprocessing pipeline began with normalizing the pixel values to ensure consistent brightness and contrast across all images. This step was crucial because poor lighting could make a happy expression appear sad, or vice versa. We applied histogram equalization to enhance the visibility of facial features, making subtle expressions more apparent to our system.

To make our system more robust, we employed data augmentation techniques that simulate natural variations in how

faces appear in real-world scenarios. We applied slight rotations, horizontal flips, and minor position shifts to create additional training examples. This approach helped our system learn that a slightly tilted head or a face positioned differently in the frame still represents the same emotion.

The feature extraction process used a ResNet-18 architecture that we adapted specifically for facial expression recognition. This network learned to identify patterns in facial features that correspond to different emotions, from obvious expressions like wide smiles to subtle cues like slight eye movements. The final output was a 512-dimensional feature vector that captured the essential characteristics of each facial expression.

4.3 Extracting Emotion from Speech Audio

The RAVDESS dataset provided us with high-quality emotional speech recordings, but converting audio signals into features that represent emotional content required sophisticated signal processing techniques. Each 3-second audio clip contained a wealth of information about how emotions manifest in human speech patterns.

We began by standardizing the audio properties across all samples. This involved normalizing volume levels so that a quietly spoken sad statement wouldn't be mistaken for a different emotion simply due to low volume. We also removed silent periods at the beginning and end of each recording to ensure our system focused on the actual speech content rather than empty space.

The transformation of audio signals into usable features involved creating mel-spectrograms - visual representations of how sound frequencies change over time. These spectrograms reveal patterns in speech that correspond to emotional expression, such as the higher frequencies associated with excitement or the lower, more monotone patterns of sadness.

From these spectrograms, we extracted Mel-Frequency Cepstral Coefficients (MFCCs), which are particularly good at capturing the spectral characteristics that humans use to perceive emotions in speech. We computed statistical measures like mean and variance across time to capture how these characteristics evolve throughout each speech sample, resulting in a 52-dimensional feature vector for each audio clip.

4.4 Our Attention Mechanism: Learning What Matters Most

4.5 Modality-Specific Processing Approach

The core of our system lies in its attention mechanisms, which we designed to work effectively with each of the three types of emotional data. The challenge was creating attention systems that could handle the very different characteristics of text embeddings, facial features, and speech characteristics while applying consistent attention principles.

We addressed this by implementing separate attention mechanisms for each modality, with each system projecting features into appropriate dimensional spaces for optimal attention computation. For text, we work with sequential BERT embeddings; for facial expressions, we process extracted ResNet features; and for speech, we apply attention to MFCC feature vectors.

4.6 Computing Attention Weights

Our attention mechanisms follow a principled approach to determining which features are most important for emotion recognition within each modality. For each input sample, we compute three key components: queries, keys, and values. These components work together to create systems that can focus on the most emotionally relevant aspects of each type of input.

The attention score computation uses the scaled dot-product attention formula, adapted specifically for emotion recognition tasks in each domain. The process calculates how compatible different features are with each other, essentially asking "which features work together to indicate a particular emotion?" The scaling factor prevents numerical instability while the softmax normalization ensures that attention weights form a proper probability distribution.

The final weighted representation combines all input features according to their emotional relevance, creating a context vector that emphasizes the most important information for emotion classification. This approach allows our systems to automatically learn what to pay attention to within each modality, rather than requiring manual specification of important features.

4.7 Classification and Training Strategy

Converting Attention-Weighted Features to Emotion Predictions

The attention-weighted features from each modality feed into carefully designed classification networks that make the final emotion predictions. We structured these networks with progressively smaller layers to gradually refine the feature representations until they can be mapped to specific emotions.

For text: 768 → 256 → 128 → 64 → 7 emotions For facial expressions: 512 → 256 → 128 → 64 → 7 emotions

For speech: 52 → 128 → 64 → 32 → 7 emotions

Between each layer, we applied dropout regularization to prevent the models from memorizing specific examples rather than learning general patterns. Batch normalization ensures stable training across different types of input data.

The final layers produce probability distributions over our seven emotion categories, allowing each system to express confidence in its predictions and handle ambiguous cases where multiple emotions might be present.

4.8 Training Process and Optimization

Our training approach balances effectiveness with computational efficiency for each modality. We use cross-entropy loss, which is well-suited for multi-class classification problems like emotion recognition. The Adam optimizer with adaptive learning rate scheduling helps the models converge efficiently while avoiding common training pitfalls.

We implemented early stopping based on validation performance to prevent overfitting - a common problem when working with datasets of different sizes and characteristics. This ensures our models generalize well to new, unseen emotional expressions rather than just memorizing the training data.

4.9 Algorithm: Attention-Based Emotion Recognition

INPUT: Sample S (can be text, facial image, or speech audio)

OUTPUT: Emotion probabilities E for seven emotions

BEGIN

Step 1: Determine input type and extract features
IF input is text sample THEN
Clean the text by removing unnecessary elements but keep emotional indicators
Break text into individual word tokens
Convert tokens to BERT embeddings creating feature matrix of size [sequence_length, 768]
Step 1: Determine input type and extract features
IF input is text sample THEN
Clean the text by removing unnecessary elements but keep emotional indicators
Break text into individual word tokens
Convert tokens to BERT embeddings creating feature matrix of size [sequence_length, 768]
ELSE IF input is facial image THEN
Normalize image brightness and contrast for consistent processing
Apply data augmentation like rotation and flipping for robustness
Extract facial features using ResNet-18 creating feature vector of size [512]
ELSE IF input is speech audio THEN
Normalize audio volume and remove silent periods
Convert audio to mel-spectrogram representation
Extract MFCC features creating feature vector of size [52]
END IF
Step 2: Apply attention mechanism to focus on important features
Create query, key, and value matrices from extracted features
Calculate attention scores by measuring compatibility between queries and keys
Apply softmax to normalize attention scores into probabilities
Compute context vector by combining values weighted by attention scores
Step 3: Classify emotions using neural network layers
Pass context vector through first dense layer with ReLU activation
Apply dropout to prevent overfitting
Apply batch normalization for stable training
Repeat the above three operations for additional hidden layers reducing dimensions progressively (256 → 128 → 64)
Pass final hidden layer through output layer with 7 neurons
Apply softmax activation to get probability distribution over emotions
RETURN emotion probabilities E
END

V.RESULT AND DISCUSSION

The evaluation of our attention-based emotion recognition methodology against established benchmarks reveals significant performance gaps across all three modalities. Contemporary research in emotion recognition has established clear performance standards that serve as evaluation criteria for new approaches. For speech emotion recognition using the RAVDESS dataset, state-of-the-art methodologies typically achieve accuracy rates between 91-98%, with good performing systems generally reaching 67-92% accuracy. These benchmarks represent the culmination of years of research employing sophisticated neural architectures, advanced feature extraction techniques, and optimized training procedures.

Table 2 Accuracy calculation

Iteration	Attention_Text	Attention_Speech	Attention_Face
1	0.32	0.41	0.27
2	0.28	0.36	0.36
3	0.4	0.33	0.27
4	0.35	0.37	0.28
5	0.31	0.39	0.3
6	0.29	0.45	0.26
7	0.33	0.34	0.33
8	0.27	0.42	0.31
9	0.36	0.31	0.33
10	0.3	0.43	0.27
Average Accuracy	0.321	0.381	0.298

Facial expression recognition on the FER2013 dataset presents similarly high performance expectations, with leading approaches achieving up to 94.2% accuracy on standard benchmark evaluations. The typical performance range for competent systems falls between 70-85% accuracy, reflecting the inherent challenges of processing facial imagery including variations in lighting, pose, and individual expression patterns. Text-based emotion recognition, while perhaps the most variable due to linguistic complexity and contextual dependencies, generally expects performance levels between 70-85% accuracy for multi-class emotion classification tasks.

5.1 Performance Gap Analysis

Our experimental results demonstrate substantial underperformance compared to these established benchmarks. The speech emotion recognition component achieved 38.1% accuracy, representing a significant deviation from the expected 67-98% range. This performance gap suggests fundamental limitations in either the feature extraction process, model architecture, or training methodology. Similarly, facial expression recognition yielded 29.8% accuracy, falling well below the anticipated 70-94% performance range. The text emotion recognition component performed at 32.1% accuracy, considerably lower than the expected 70-85% range.

These performance disparities indicate that while attention mechanisms offer theoretical advantages for emotion recognition tasks, their standalone application in our current implementation may be insufficient for achieving competitive results. The consistent underperformance across all modalities suggests systematic issues rather than modality-specific problems, pointing toward fundamental architectural or methodological limitations that require addressing.

VI.CONCLUSION AND FUTURE SCOPE

6.1 Research Summary and Key Findings

This research investigated the effectiveness of attention mechanisms in emotion recognition across three distinct modalities: textual expressions, facial imagery, and vocal patterns. Through the implementation of separate attention-enhanced models for each modality, we aimed to understand which types of emotional data benefit most from attention-based processing and how these mechanisms compare in their ability to identify emotionally relevant features.

Our experimental results reveal significant insights into the comparative performance of attention mechanisms across different emotional modalities. The speech emotion recognition component achieved the highest average accuracy of 38.1%, followed by text-based emotion recognition at 32.1%, and facial expression recognition at 29.8%. While these results demonstrate the feasibility of applying attention mechanisms to emotion recognition tasks, they also highlight substantial performance gaps when compared to established benchmarks in the field.

The findings indicate that attention mechanisms show differential effectiveness across modalities, with speech data benefiting most from attention-based processing. This suggests that the temporal and spectral patterns inherent in vocal emotional expression are particularly well-suited to attention-based feature selection. However, the overall performance levels across all modalities remain significantly below contemporary standards, indicating fundamental limitations in the current approach.

6.2 Limitations and Contributing Factors

Several factors likely contribute to the observed performance gaps. Dataset size limitations may prevent the attention mechanisms from learning sufficiently complex patterns necessary for accurate emotion classification. The relatively simple feature extraction approaches, particularly the use of basic MFCC features for speech analysis, may not capture the rich spectral and temporal characteristics that modern high-performing systems exploit. Additionally, the standalone attention architecture may lack the representational capacity that hybrid approaches combining attention with convolutional or recurrent components typically provide.

Training strategy limitations, including potentially insufficient training epochs, suboptimal hyperparameter configurations, or inadequate optimization procedures, could further constrain performance. The preprocessing pipelines, while designed to preserve emotional content, may not adequately address the specific requirements of each modality or properly normalize the data for optimal attention mechanism operation. Furthermore, the absence of cross-modal validation and the lack of ensemble approaches may have limited the system's ability to leverage complementary information across different emotional expression channels.

The computational constraints and limited exploration of attention mechanism variants also represent significant limitations. The current implementation focuses on basic self-attention mechanisms without investigating more sophisticated attention architectures such as multi-head attention, hierarchical attention, or transformer-based approaches that have proven successful in other domains.

6.3 Research Contributions and Implications

Despite the performance limitations, this research makes several important contributions to the field of emotion recognition. The systematic comparison of attention mechanisms across three distinct modalities provides valuable insights into the differential effectiveness of attention-based approaches for various types of emotional data. The finding that speech emotion recognition benefits most from attention mechanisms offers practical guidance for future system development and resource allocation.

The research also demonstrates the importance of careful benchmarking and performance evaluation in emotion recognition research. By comparing our results against established standards, we provide a realistic assessment of where attention-based approaches currently stand and what improvements are necessary for practical deployment. This transparency in reporting both successes and limitations contributes to the broader understanding of attention mechanisms in emotion recognition contexts.

Additionally, the unified methodological framework developed for this research provides a foundation for future comparative studies across different emotion recognition modalities. The systematic approach to feature extraction, attention implementation, and performance evaluation can serve as a template for researchers investigating similar cross-modal emotion recognition questions.

6.4 Future Research Directions

6.4.1 Architectural Enhancements

Addressing the identified performance limitations requires a multi-faceted approach focusing on architectural improvements, enhanced feature extraction, and optimized training procedures. Implementing ensemble methods that combine attention mechanisms with proven architectures such as convolutional neural networks and long short-term memory networks could provide the representational diversity necessary for improved performance. Advanced feature extraction techniques, including sophisticated spectral analysis for speech and deep convolutional features for facial imagery, would likely enhance the quality of input representations.

Future research should explore transformer-based architectures specifically designed for emotion recognition tasks, incorporating multi-head attention mechanisms and positional encoding schemes adapted for temporal and spatial emotional data. The investigation of hierarchical attention models that can capture both local and global emotional patterns represents another promising direction for architectural innovation.

6.4.2 Data and Training Improvements

Data augmentation strategies could effectively expand the training datasets, potentially addressing the dataset size limitations that may constrain learning. Transfer learning approaches, utilizing pre-trained models as feature extractors, could provide more robust initial representations while reducing training requirements. Systematic hyperparameter optimization, including careful tuning of learning rates, batch sizes, and network architecture parameters, would likely yield performance improvements across all modalities.

Future work should also investigate advanced data preprocessing techniques, including sophisticated normalization approaches for cross-modal data integration and novel augmentation strategies

that preserve emotional content while increasing data diversity. The development of larger, more diverse datasets specifically designed for attention-based emotion recognition would provide crucial resources for model development and evaluation.

6.4.3 Multi-Modal Integration

While this research focused on individual modality analysis, future investigations should explore true multi-modal fusion approaches that can leverage the complementary information available across different emotional expression channels. The development of attention mechanisms that can simultaneously process textual, visual, and auditory emotional cues represents a significant opportunity for performance improvement.

Cross-modal attention mechanisms that can identify relationships between different types of emotional expression could provide insights into how humans naturally integrate multiple sources of emotional information. Such approaches might reveal previously unknown patterns in emotional expression and lead to more robust emotion recognition systems.

6.4.4 Real-World Applications and Deployment

Future research should focus on developing practical applications of attention-based emotion recognition systems, including real-time processing capabilities and deployment in resource-constrained environments. The investigation of edge computing approaches for emotion recognition could expand the practical applicability of these systems to mobile and embedded applications.

The development of interpretable attention mechanisms that can provide insights into their decision-making processes would enhance the trustworthiness and adoption of emotion recognition systems in sensitive applications such as healthcare, education, and human-computer interaction.

VII.REFERENCES

1. Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. *Entropy*. https://doi.org/10.3390/e25101440
2. *Enhancing Speech Emotion Recognition Through Advanced Feature Extraction and Deep Learning: A Study Using the RAVDESS Dataset*. (2024). 1–7. https://doi.org/10.1109/iciteics61368.2024.10624849
3. Bhangale, K. B., & Kothandaraman, M. (2023). Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics*, *12*(4), 839. https://doi.org/10.3390/electronics12040839
4. Speech Emotion Classification: A Survey of the State-of-the-Art. (2023). *Lecture Notes in Computer Science*, 379–394. https://doi.org/10.1007/978-3-031-25271-6_24
5. Avabratha, V., Rana, S., Narayan, S., Raju, S. Y., & Sahana, S. (2024). *Speech and Facial Emotion Recognition using Convolutional Neural Network and Random Forest: A Multimodal Analysis*. 1–5. https://doi.org/10.1109/apcit62007.2024.10673495
6. K, A. R. (2024). Advancements in Emotion Detection: A Comprehensive Review of Text and Audio-Based Approaches. *International Journal For Science Technology And Engineering*, *12*(3), 2875–2879. https://doi.org/10.22214/ijraset.2024.59451
7. Abdykerimova, L., Abdikerimova, G., Konyrkhanova, A., Nurova, G., Bazarova, M., Bersugir, M., Kaldarova, M., & Yerzhanova, A. Ye. (2024). Analysis of the emotional coloring of text using machine and deep learning methods. *International Journal of Electrical and Computer Engineering*. https://doi.org/10.11591/ijece.v14i3.pp3055-3063
8. Shah, H., Shah, H., & Chopade, M. (2024). *Text-Based Emotion Recognition: A Review*. 551–561. https://doi.org/10.1007/978-981-99-7954-7_49
9. Chaithra, I. V., & Samanvaya, K. J. (2024). *Text-Based Emotion Recognition Using Deep Learning*. 1–7. https://doi.org/10.1109/icait61638.2024.10690782
10. Asghar, M., Lajis, A., Alam, M. M., Rahmat, Mohd. K., Nasir, H., Ahmad, H., Al-Rakhami, M., Alamri, A., & Albogamy, F. (2022). A Deep Neural Network Model for the Detection and Classification of Emotions from Textual Content. *Complexity*, *2022*, 8221121:1-8221121:12. https://doi.org/10.1155/2022/8221121
11. Drakopoulos, G., Pikramenos, G., Spyrou, E. D., & Perantonis, S. (2019). Emotion Recognition from Speech: A Survey. *International Conference on Web Information Systems and Technologies*, *2*, 432–439. https://doi.org/10.5220/0008495004320439
12. Kashif, K., Hayat, S., & Khan, M. E. (2016). Emotion Detection through Text: Survey. *TIJ's Research Journal of Science & IT Management - RJSITM*, *5*(7). https://www.theinternationaljournal.org/ojs/index.php?journal=rjitsm&page=article&op=view&path%5B%5D=4951
13. Deng, J., & Ren, F. (2021). A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Transactions on Affective Computing*, *01*, 1. https://doi.org/10.1109/TAFFC.2021.3053275
14. Seyeditabari, A., Tabari, N., & Zadrozny, W. (n.d.). *Emotion Detection in Text: a Review*. https://doi.org/10.48550/arxiv.1806.00674
15. Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-Emotion Neural Network for Emotion Recognition From Text. *IEEE Access*, *7*, 111866–111878. https://doi.org/10.1109/ACCESS.2019.2934529
16. Jain, L. (2024). Face Emotion Recognition (FER) Using Convolutional Neural Network (CNN) in Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2024.58077
17. Gupta, P. K., Varadharajan, N., Ajith, K., Singh, T., & Patra, P. (2024). *Facial Emotion Recognition Using Computer Vision Techniques*. 1–7. https://doi.org/10.1109/icccnt61001.2024.10725699
18. Mahastama, A. W., Mahendra, E., Chrismanto, A. R., Rini, M. N. A., & Prabawati, A. G. (2024). Facial Expression Classification System Using Stacked CNN. *International Journal of Advanced Computer Science and Applications*, *15*(10). https://doi.org/10.14569/ijacsa.2024.0151049
19. Khan, A. (2022). Facial Emotion Recognition Using

Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. *Information*, *13*(6), 268. https://doi.org/10.3390/info13060268

20. Singh, S., & Shukla, S. (2023). *Deep Learning Approach to Emotion Recognition by Facial Expressions: A Review Paper*. https://doi.org/10.1109/icac3n60023.2023.10541462