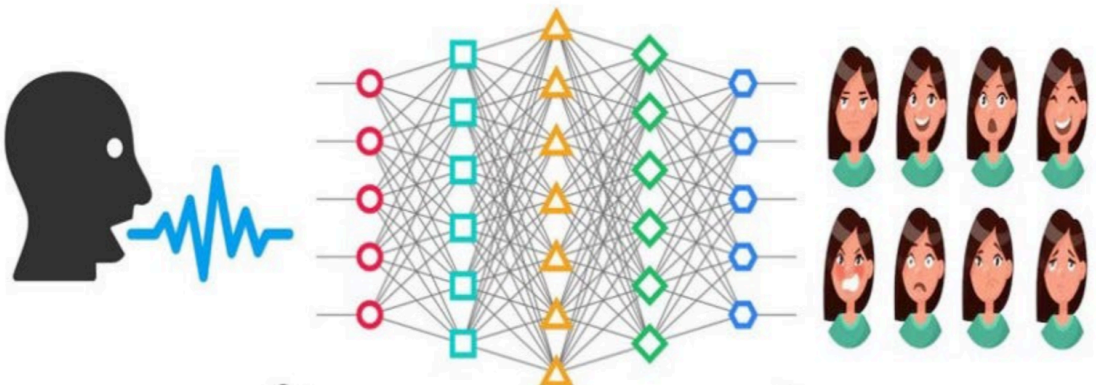
काशी हिन्दू विश्वविद्यालय

BANARAS HINDU UNIVERSITY

MINOR DISSERTATION

ON

EMOTION RECOGNITION WITH MULTI MODAL DATA USING AI TECHNIQUES



Guided by

Mentor - Prof. Manjari Gupta

Scholar - Mr. Bethany Gosala

Name-

Anshika Singh - 23419CAS004

Semester - III

M.Sc. Computational Sciences & Applications

DST-CIMS

Banaras Hindu University

T A B L E O F C O N T E N T S

1. A B S T R A C T

Emotion recognition with multimodal data has gained significant attention due to its applicability in areas such as human-computer interaction, healthcare, and entertainment. By combining audio, visual, and textual inputs, multimodal emotion recognition achieves higher accuracy than unimodal approaches. This report provides an overview of multimodal emotion recognition using AI techniques, focusing on Gaurav Sahu's research, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," and the IEMOCAP dataset. The study demonstrates how combining modalities enhances emotion detection accuracy and resolves ambiguities in emotional expressions. The report discusses existing literature, methodology, challenges, and future research directions to improve emotion recognition systems.

2. I N T R O D U C T I O N

2.1 Emotion Recognition

Emotion recognition is a crucial area of research in human-computer interaction, aiming to bridge the gap between machines and human emotions. Understanding emotions enhances the ability of AI systems to interact in a more natural and empathetic manner, enabling applications in mental health monitoring, customer service, education, and entertainment. Traditional emotion recognition methods focus on a single modality, such as analyzing textual sentiment or acoustic features of speech. While effective in specific scenarios, these unimodal approaches often struggle to accurately interpret emotions due to the inherent complexity and variability of human expression. For instance, sarcasm or subtle emotional cues may be lost when relying solely on text, while speech alone may not fully capture the context of an emotional state.

AI techniques, especially deep learning, have revolutionized emotion recognition by enabling systems to process and learn from vast and diverse datasets. Models like **recurrent neural networks (RNNs)** and **transformers** are adept at analyzing sequential and contextual information, making them ideal for tasks like sentiment analysis and speech emotion detection. However, even the most advanced unimodal systems face challenges in resolving ambiguities and achieving human-level understanding of emotions. This limitation has led researchers to explore multimodal approaches that integrate information from multiple data sources, providing a richer and more holistic view of emotional states.

Multimodal data, which combines inputs from various modalities such as text, speech, and visual cues, offers a powerful solution to the challenges of emotion recognition. Multimodal systems capitalize on the strengths of each modality to address the limitations of others. For example, speech carries prosodic features like pitch, tone, and rhythm, which are critical for detecting emotional intensity, while text provides semantic context that clarifies ambiguous vocal expressions. By fusing these modalities, multimodal systems achieve higher accuracy and robustness in emotion detection compared to their unimodal counterparts.

The use of multimodal data also introduces challenges, such as the need for effective feature fusion techniques and the management of varying data formats and temporal alignment. Advanced AI models, including **multimodal transformers** and **late-fusion architectures** [8], have been developed to address these issues. These approaches enable models to learn complementary features from each modality while preserving their unique characteristics. The integration of multimodal data has already shown promise in applications like virtual assistants, emotion-aware AI systems, and mental health diagnostics, positioning it as a transformative advancement in the field of emotion recognition.

2.2 Multimodal Data

Multimodal emotion recognition leverages data from multiple sources to provide a comprehensive analysis of emotional states, addressing the complexity and variability of human expression. Each modality captures unique emotional signals, and their integration offers a more robust understanding. For instance, **audio** captures tonal and prosodic features like pitch, rhythm, and speech intensity, which are pivotal for detecting emotions like anger or happiness. **Text** provides semantic and contextual information, decoding emotions embedded in words, phrases, and sentence structures. **Visual cues**, such as facial expressions, gestures, and body posture, enrich the analysis by reflecting non-verbal emotional nuances often absent in audio or text.

The integration of these modalities enhances emotion recognition by compensating for the limitations of single-modal systems. Key advantages and challenges of multimodal approaches include:

- **Rich Feature Representation**: Combining audio, text, and visual data offers a multidimensional perspective of emotions, crucial for detecting subtle or conflicting emotional states.
- **Improved Accuracy**: Multimodal systems outperform unimodal approaches by leveraging the complementary strengths of each modality.
- **Temporal Synchronization**: Aligning modalities that occur simultaneously, such as speech and facial expressions, requires advanced techniques like cross-modal attention mechanisms.
- **Fusion Techniques**: Effective fusion methods, such as early, late, or hybrid fusion, are critical for integrating heterogeneous data while preserving the unique characteristics of each modality.

By addressing these challenges, multimodal emotion recognition has become a transformative tool in applications like mental health diagnostics, empathetic AI systems, and human-computer interaction.

2.3 IEMOCAP Dataset Overview

The IEMOCAP dataset (Interactive Emotional Dyadic Motion Capture) is a widely used resource for multimodal emotion analysis. It contains 12 hours of dyadic interactions, annotated for emotions such as anger, happiness, sadness, and neutrality. The dataset includes:

- Audio: Speech segments labeled for emotions.
- Video: Motion-tracked facial expressions and gestures.
- Text: Transcriptions of dialogues.
  Its comprehensive multimodal annotations make it ideal for developing AI-driven emotion recognition systems.

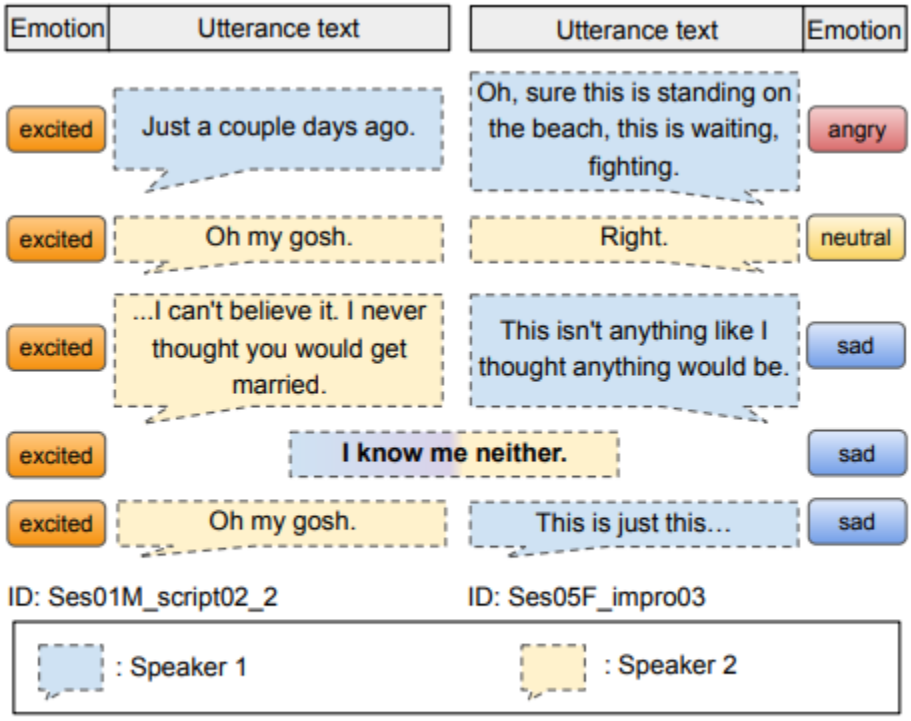


Figure 1: Examples of temporal effects on conversations

3. RELATED WORKS

In the field of multimodal emotion recognition, numerous studies have explored the integration of multiple data modalities to improve the accuracy and robustness of emotion detection systems. These works illustrate the potential of combining audio, text, and visual features for a holistic understanding of emotional states.

3.1 Integration of Audio and Text for Emotion Detection

Poria et al. [1] proposed a multimodal emotion recognition framework that combines acoustic and textual features. They utilized a hierarchical fusion approach to integrate audio and text modalities, achieving significant improvements in emotion detection accuracy. Their work demonstrated the importance of leveraging complementary features from multiple modalities for robust emotion recognition systems.

3.2 Attention Mechanisms for Multimodal Emotion Analysis

Zadeh et al. [1] introduced the Multimodal Transformer (MulT), which uses cross-modal attention mechanisms to align and fuse audio, text, and visual data. Their model achieved state-of-the-art performance on emotion recognition tasks by learning intricate relationships between modalities, highlighting the effectiveness of attention-based architectures in multimodal systems.

3.3 Deep Learning for Speech and Text Emotion Recognition

Akçay et al. [2] developed a deep learning framework that integrates CNNs for speech feature extraction and RNNs for text analysis. By combining the temporal dynamics of speech with semantic information from text, their model outperformed unimodal systems in detecting complex emotional states.

3.4 Temporal Alignment in Multimodal Systems

Tsai et al. [3] addressed the challenge of temporal synchronization in multimodal data by proposing a sequence-to-sequence model that aligns audio and text streams at a fine-grained temporal level. Their method significantly improved the performance of multimodal emotion recognition tasks, especially in scenarios with asynchronous data streams.

3.5 Feature Fusion Techniques for Multimodal Data

Baltrušaitis et al. [4] provided a comprehensive review of feature fusion strategies for multimodal emotion recognition. They categorized fusion approaches into early, late, and hybrid

fusion and discussed their trade-offs. Their work serves as a valuable resource for researchers exploring feature integration techniques in multimodal systems.

Building on these foundational works, the current study focuses on integrating audio and text modalities using advanced deep learning architectures to improve emotion recognition accuracy. By addressing challenges like feature fusion, temporal alignment, and modality-specific noise, this research aims to contribute to the development of more robust and generalized multimodal emotion recognition frameworks.

4. LITERATURE REVIEW

4.1 Multimodal Approaches

Gaurav Sahu's research, *"Multimodal Speech Emotion Recognition and Ambiguity Resolution,"* highlights the importance of integrating multiple data modalities to enhance emotion recognition. Multimodal fusion leverages the strengths of each modality—audio, video, and text—to provide a more comprehensive understanding of emotional expressions. By addressing the limitations of unimodal systems, multimodal approaches reduce ambiguities, such as distinguishing between closely related emotions like neutrality and mixed feelings. For instance, subtle variations in vocal tone that might be unclear on their own can be clarified when combined with facial expressions or contextual text.

Fusion strategies play a central role in this integration:

- *Early Fusion*: Combines raw features from all modalities into a unified representation before classification.
- *Intermediate Fusion*: Merges modalities during model training, allowing each to retain specific features while interacting with others.
- *Late Fusion*: Integrates predictions from individual modalities at the decision level, preserving modality-specific information.

4.2 AI Models

Recent advancements in AI have introduced powerful architectures for processing multimodal data:

- **Convolutional Neural Networks (CNNs)**: Extract patterns from visual features in video data, such as facial landmarks or expressions.
- **Recurrent Neural Networks (RNNs)**, including LSTMs and GRUs: Capture sequential dependencies in audio data like tone and pitch variations.
- **Transformer-based Models**: Analyze text for semantic and contextual understanding, particularly effective for capturing nuanced emotional cues.

By combining these models, multimodal systems achieve enhanced reliability and accuracy in emotion classification. This layered approach enables complementary modalities to address inherent ambiguities and improve system robustness.

5. M A T E R I A L & M E T H O D S

5.1 Computational Requirements for Experiment

The experiments in multimodal emotion recognition demand substantial computational resources due to the complexity of processing and integrating data from multiple modalities. The training and evaluation of deep learning models, such as transformers and convolutional neural networks (CNNs), require high-performance hardware capable of handling large datasets and high-dimensional feature spaces. Specifically, **GPUs with CUDA support** are essential for accelerating matrix operations and parallel computations. Additionally, sufficient **RAM (16GB or more)** and **storage** are necessary to manage and preprocess multimodal datasets, which often include audio waveforms, text transcripts, and auxiliary metadata. Software frameworks like **PyTorch**, **TensorFlow**, and libraries for natural language processing (e.g., Hugging Face Transformers) and audio processing (e.g., Librosa) are integral to the experimental pipeline. These requirements ensure efficient handling of data and optimal performance during training and testing phases.

5.2 Dataset Description

The **IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset** is a widely used benchmark for emotion recognition tasks. It consists of multimodal data collected from dyadic conversations performed by 10 actors over five sessions. The dataset captures emotional expressions through **audio**, **text**, and **visual modalities**, making it ideal for studying multimodal emotion recognition. Each session includes scripted and improvised dialogues designed to elicit a range of emotions such as anger, happiness, sadness, and neutral states. The annotations are provided at both **utterance level** and **session level**, enabling detailed analysis of emotion dynamics.

The dataset includes over **12 hours of audiovisual recordings**, segmented into more than **10,000 utterances**, with each labeled for one of the predefined emotional categories. The audio data consists of speech recordings, while the text data provides transcriptions of the dialogues. While the visual modality includes facial expression data, this study focuses on the **audio and text modalities** for emotion recognition. The rich multimodal nature of IEMOCAP makes it a valuable resource for exploring the integration of heterogeneous data sources in emotion recognition tasks.

Key Features of the IEMOCAP Dataset:

- **Modalities**: Audio, text, and visual data.

- **Annotations**: Emotion labels at utterance and session levels.
- **Emotions**: Categories include anger, happiness, sadness, neutral, and others.
- **Data Volume**: Over 12 hours of recordings and 10,000+ utterances.
- **Format**: Audio in WAV files, transcriptions in text format, and facial data in motion capture files.
- **Language**: English, providing a standardized benchmark for emotion recognition tasks.

By leveraging the IEMOCAP dataset, this study aims to analyze the complementary roles of audio and text modalities in enhancing the accuracy of emotion detection systems.

5.3 Data Pre-Processing

The code follows a multimodal approach with separate models for processing audio and text data, then combines them into a unified model for final classification. Here's a breakdown of the architecture:

Audio Model

- **Convolutional Neural Network (CNN)**:
    - Convolutional layers are used to process the audio features extracted earlier. CNNs help identify patterns within these features, such as changes in pitch or tone that correlate with emotional expressions.
    - **Pooling Layers**: Max pooling or average pooling layers reduce the spatial dimensions, making the model computationally efficient and focusing on the most relevant features.
    - **Dense Layers**: After the convolutional and pooling layers, dense (fully connected) layers are added. These layers help in learning more abstract patterns from the convolutions, transforming the features for combination with the text model.

Text Model

- **Recurrent Neural Network (RNN)**:
    - The text model often uses RNNs (such as LSTMs or GRUs) to capture sequential information from the text data. These are useful for understanding the flow and emphasis within sentences, both of which contribute to emotional tone.
    - **Dense Layers**: Similar to the audio model, dense layers follow the RNN layers to help map the RNN's outputs to a feature space that aligns with the audio features for better fusion.

Fusion Layer

- **Concatenation**:
    - After processing the audio and text data independently, the feature outputs from each are concatenated. This merged feature set allows the model to consider both modalities jointly when making its prediction.
- **Final Dense Layers**: A series of dense layers follow the concatenation to refine the combined feature representation and prepare it for classification.
- **Output Layer**: The final layer uses softmax (or another activation function suitable for classification) to output probabilities for each emotion class.

5.4 Fusion Strategies

5.4.1 Intermediate Fusion

Intermediate fusion leverages attention-based mechanisms to integrate features from different modalities during the training process. By aligning audio and text features at an intermediate layer, the model ensures that each modality's unique characteristics are preserved while enabling interactions between them. This approach allows the model to emphasize critical aspects of each modality based on context, such as prioritizing vocal tone for spoken emotions or textual cues for nuanced expressions. For example, the model can align an emotionally charged word in text with corresponding pitch changes in audio.

Key Features of Intermediate Fusion:

- Preserves modality-specific features while enabling cross-modality interactions.
- Uses attention mechanisms to highlight relevant features from each modality.
- Allows dynamic weighting of modalities during training, improving alignment and accuracy.

5.4.2 Decision-Level Fusion

In decision-level fusion, predictions from individual modality-specific models are combined to make a final classification. Each modality contributes its strengths independently, and their outputs are weighted to optimize overall accuracy. This method ensures that modality-specific errors do not overly influence the final prediction, making the system robust to noise or missing data in any single modality. For instance, if audio data is unclear, the system can rely more on text features for classification.

Key Features of Decision-Level Fusion:

- Combines independent predictions from modality-specific models.

- Uses weighted averaging or voting schemes to balance contributions from each modality.
- Enhances robustness by mitigating the impact of noisy or incomplete data.

These fusion strategies complement each other, with intermediate fusion enabling seamless feature integration during training and decision-level fusion ensuring robust predictions at inference. Together, they maximize the strengths of multimodal systems and improve emotion recognition reliability.

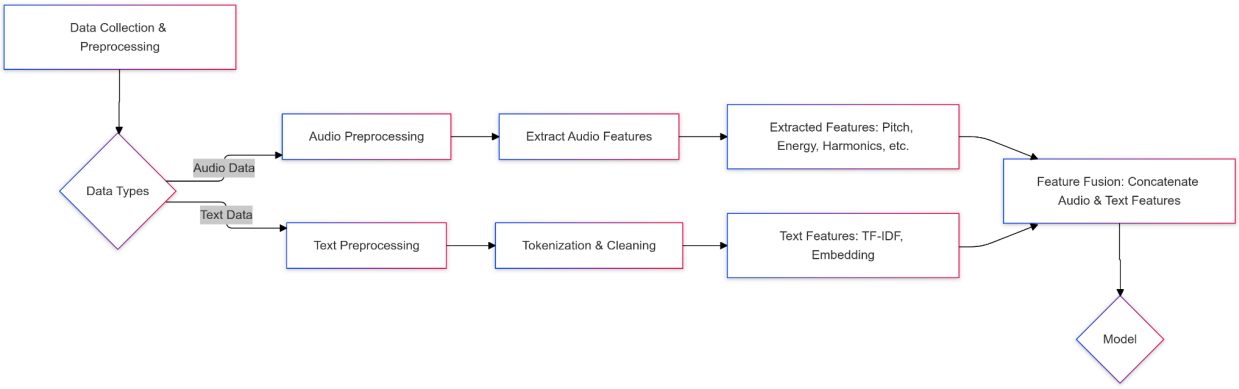


Figure 2(a): Flowchart of steps from data collection to model building

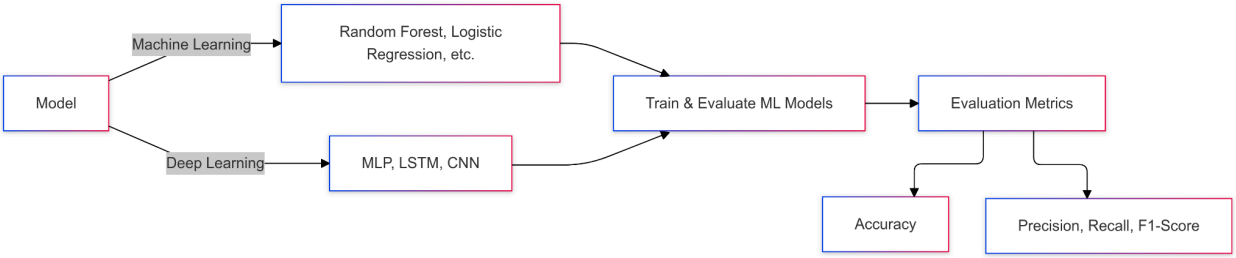


Figure 2(b): Flowchart of model building

5.5 Models

The multimodal emotion recognition system utilized a combination of machine learning (ML) and deep learning (DL) models tailored to extract and process features from different modalities. Each model was designed to capture the unique characteristics of its respective modality, and their outputs were fused for joint analysis and classification.

5.5.1 Machine Learning Models

- **Support Vector Machines (SVMs)**:
  SVMs were used as baseline models for unimodal classification tasks, particularly with audio features like MFCCs. Their ability to find optimal decision boundaries in high-dimensional spaces made them effective for initial experimentation.
- **Random Forests**:
  Random Forest classifiers were employed for feature selection and initial classification tasks in unimodal setups. Their robustness to overfitting and ease of interpretability provided valuable insights into the importance of different features.

5.5.2 Deep Learning Models

- **Convolutional Neural Networks (CNNs)**:
  CNNs were applied to extract spatial features from audio data (spectrograms) and visual data (facial expressions). Their hierarchical feature extraction capability allowed the model to identify patterns and emotional cues effectively.
    - *Audio*: CNNs processed MFCCs and pitch features to capture temporal variations in emotional tone.
    - *Video*: CNNs analyzed facial landmarks and motion-tracked gestures to extract features indicative of emotions.
- **Recurrent Neural Networks (RNNs)**:
  RNNs, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, were used for text analysis to capture the sequential nature of dialogue. These models excelled at understanding the flow and emphasis in sentences, which are critical for emotional context.
- **Transformer-based Models**:
  Pre-trained transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), were utilized for text embeddings. These models provided a deep contextual understanding of text, capturing nuanced emotional expressions in dialogues.

```
Model,Accuracy,F-score,Precision,Recall
Random forest,0.6602040816326531,0.6589746509577581,0.710318573043804,0.6551234252953043
CNN,0.6678571428571428,0.6758247039226336,0.6847388819941825,0.67949569593711
CNN_20,0.6622448979591836,0.6789123107303124,0.675315220282427,0.685257742354778
Transformer,0.6377551020408163,0.6496377822161042,0.6416410189219881,0.6653526690721016
CNN_Improved,0.29744897959183675,0.07641892777559313,0.04957482993197279,0.16666666666666666
improved_CNN,0.6545918367346939,0.6619958892425695,0.6583469763380236,0.6796455303250403
CNN_Optuna,0.6602040816326531,0.6635684197631221,0.675151109221947,0.6693046978149089
```

5.6 Evaluation Metrics

- Evaluation was performed using standard metrics: **accuracy**, **precision**, **recall**, and **F1-score**. Cross-validation was used to ensure that the model generalized well across different sessions in the dataset.

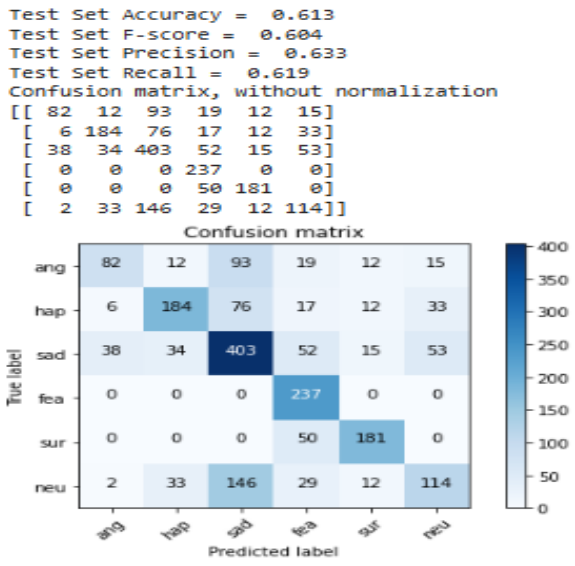


Figure 3: **Random forest model** for **sentence classification**

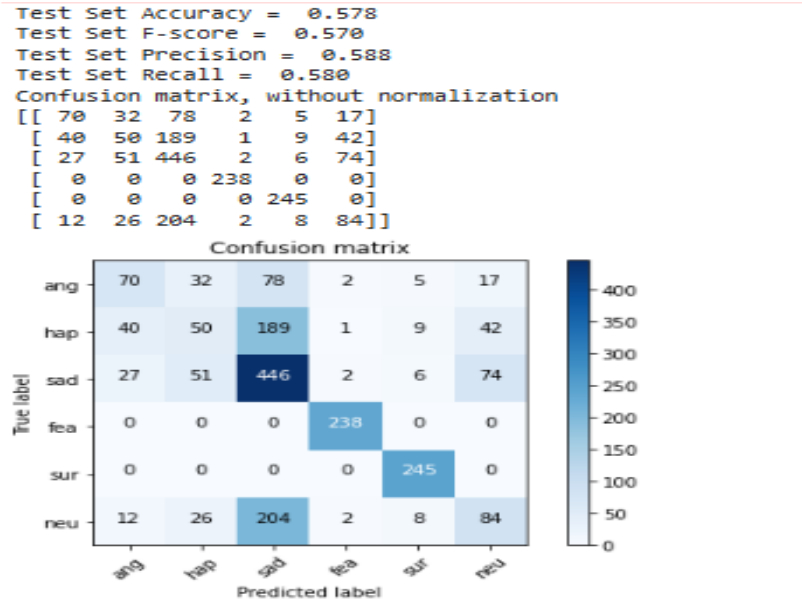


Figure 4: **Random forest model** for **audio classification**

```
Test Set Accuracy =  0.666
Test Set F-score =  0.663
Test Set Precision =  0.708
Test Set Recall =  0.658
Confusion matrix, without normalization
[[ 87  14  86   0   3  14]
 [ 23 137 131   5   5  30]
 [ 23  21 509   2   4  47]
 [  0   0   0 238   0   0]
 [  0   0   0   0 245   0]
 [ 14  24 204   1   3  90]]
```

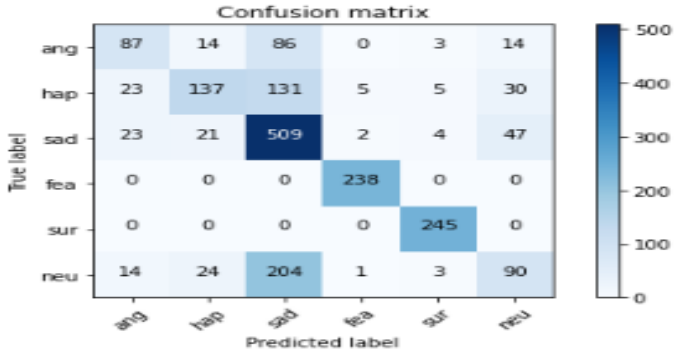


Figure 5: **Random forest model** for **multimodal combined classification**

```
Epoch 20/20
196/196 [==============================] - 105s 537ms/step - loss: 0.3678 - accuracy: 0.8625 - val_loss: 1.3002 - val_accurac
y: 0.6531
62/62 [==============================] - 5s 82ms/step
```

```python
cnn_classifier = Sequential([
    Conv1D(64, kernel_size=3, activation='relu', input_shape=(combined_x_train_cnn.shape[1], 1)),
    MaxPooling1D(pool_size=2),
    Conv1D(128, kernel_size=3, activation='relu'),
    MaxPooling1D(pool_size=2),
    Flatten(),
    Dense(256, activation='relu'),
    Dropout(0.5),
    Dense(len(emotion_dict), activation='softmax')
])
```

`Figure 6: **1D CNN model** for **combined audio-text classification**

6. C H A L L E N G E S F A C E D

6.1 Modality Synchronization

One of the critical challenges faced during the project was aligning audio, video, and text data streams, as each modality had differing start and end times. These temporal mismatches often led to misaligned features, which could disrupt the model's ability to correlate emotional cues across modalities. For instance, a vocal cue indicating anger might not align with the corresponding facial expression, causing inconsistencies in the dataset. To address this, synchronization techniques were applied, such as aligning timestamps and interpolating missing data. However, achieving perfect synchronization, especially in dynamic scenarios involving rapid or overlapping interactions, remained a challenge.

Key Issues with Modality Synchronization:

- Temporal mismatches between modalities.
- Difficulty handling overlapping or rapid interactions.
- Dependence on accurate timestamps for alignment.

6.2 Data Imbalance and Ambiguity in Emotions

The dataset exhibited an imbalance in emotional categories, with certain emotions like "neutral" and "anger" being more prevalent than others like "happiness" or "fear." This imbalance caused the model to skew toward majority classes, reducing its effectiveness in detecting underrepresented emotions. To overcome this, techniques like SMOTE (Synthetic Minority Oversampling Technique) and data augmentation were employed to create a balanced dataset.

Ambiguity in emotional expressions presented another significant challenge. Similar emotions, such as "happiness" and "excitement," or "frustration" and "anger," often overlapped in their features, making it difficult for the model to differentiate between them. Intermediate fusion strategies, which emphasized context from all modalities, helped to resolve some of these ambiguities by leveraging complementary cues like textual context or facial expressions alongside vocal tones.

Key Challenges with Data Imbalance and Ambiguity:

- Overrepresentation of certain emotional classes.
- Difficulty distinguishing subtle differences in overlapping emotions.
- Limited effectiveness of traditional augmentation techniques for nuanced emotional data.

6.3 Computational Costs

The computational demands of training multimodal models posed another significant challenge. Feature extraction, model training, and hyperparameter tuning required substantial GPU resources and memory. These high costs were particularly burdensome when experimenting with advanced architectures like transformers and attention mechanisms. To mitigate these issues, pre-trained models were utilized for feature extraction, and dimensionality reduction techniques were applied to streamline the processing pipeline without compromising accuracy.

Key Computational Challenges:

- High GPU and memory usage for multimodal training.
- Long training times for complex models.
- Balancing computational efficiency with model performance.

7. CONCLUSION

This dissertation demonstrates that multimodal data, specifically the combination of text and audio, significantly enhances the performance of emotion recognition through machine learning and deep learning models compared to individual modalities.

The findings indicate that integrating both text and audio data leads to improved accuracy. Interestingly, the study also reveals that simple machine learning models, such as Random Forests or Logistic Regression, perform comparably to more complex deep learning models, suggesting that for certain tasks, the additional complexity of deep learning may not always yield substantial improvements or we need to incorporate some additional changes to deep learning models. These results underscore the potential of multimodal fusion for improving model accuracy while also highlighting the efficiency and practicality of using simpler machine learning algorithms in such tasks.

8. FUTURE WORKS

Future Research Directions:

- **Integration of Additional Modalities**: Explore the use of physiological signals like EEG or galvanic skin response for deeper emotional analysis.
- **Advanced Fusion Techniques**: Implement attention-based and hierarchical fusion strategies to improve inter-modality communication.
- **Real-Time Applications**: Develop computationally efficient models capable of operating on edge devices for real-time emotion recognition.
- **Dataset Expansion**: Include diverse datasets with varied cultural and linguistic contexts to improve generalization.
- **Robustness Improvements**: Address challenges such as noise, missing data, and cross-domain adaptability to enhance system reliability.

By addressing these areas, future work can build on the current findings, driving innovation in multimodal emotion recognition systems and broadening their applicability in real-world scenarios.

9. REFERENCES

1. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
2. Sahu, G. (Year). Multimodal Speech Emotion Recognition and Ambiguity Resolution. David R. Cheriton School of Computer Science, University of Waterloo.
3. Praveen, R. G., de Melo, W. C., Ullah, N., Aslam, H., Zeeshan, O., Denorme, T., ... & Granger, E. (2022). A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2486-2495)
4. Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-Based Systems, 244, 108580
5. Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion, 99, 101847
6. Singh, M., & Fang, Y. (2020). Emotion recognition in audio and video using deep neural networks. arXiv preprint arXiv:2006.08129
7. https://www.kdnuggets.com/2023/03/multimodal-models-explained.html
8. https://arxiv.org/abs/2002.06305
9. Poria, S., et al. "A Deep Convolutional Framework for Audiovisual Emotion Recognition." Paper Link
10. Zadeh, A., et al. "Multimodal Transformer for Unaligned Multimodal Language Sequences." Paper Link
11. Akçay, M., and Oguz, K. "Speech Emotion Recognition: A Review of Datasets and Deep Learning Techniques." Paper Link
12. Tsai, Y.-H. H., et al. "Learning Factorized Multimodal Representations." Paper Link
13. Baltrušaitis, T., et al. "Multimodal Machine Learning: A Survey and Taxonomy." Paper Link