Emotion Recognition from Video Audio and Text

Ms. G Srilakshmi, Putta Meghana, Chakali Mythri, Kattukuri Samhitha

[1]Associate Professor, Department Of Ece, Bhoj Reddy Engineering College For Women, India.

[2,3,4]B. Tech Students, Department Of Ece, Bhoj Reddy Engineering College For Women, India.

ABSTRACT

Emotion detection from video, audio, and text has become a key focus in artificial intelligence and human-computer interaction. As digital communication increasingly involves multiple modalities, accurately interpreting human emotions is essential for enhancing user experience, supporting mental health diagnostics, and advancing affective computing. This paper provides a comprehensive overview of emotion recognition methods across video, audio, and text, exploring the unique contributions of each modality and their combined potential in multimodal systems.

The analysis begins by examining how each modality contributes to emotion detection. Video leverages facial expressions, gestures, and body language using computer vision, while audio focuses on vocal traits such as pitch and tone through signal processing and machine learning. Text-based emotion detection uses natural language processing to interpret sentiment and context from written content. When integrated, these modalities create more accurate and resilient emotion recognition systems that mirror the complexity of human emotion.

The paper also explores key challenges, including data synchronization, multimodal feature extraction, and the scarcity of diverse, annotated datasets. Advances in machine learning, especially deep learning techniques like transformers and attention mechanisms, have significantly improved emotion detection performance. Potential applications include mental health monitoring, customer service, education, and interactive entertainment. The paper concludes by highlighting future research needs, including ethical considerations, generalizable models, and real-time emotion recognition, aiming to create AI systems that better understand and respond to human emotions.

1-INTRODUCTION

In recent years, the field of artificial intelligence (AI) has seen remarkable advances, particularly in understanding and processing human emotions. Emotions are an intrinsic part of human behavior and play a critical role in communication, decision-making, and social interaction. Emotion recognition, therefore, aims to enable machines and systems to detect, interpret, and respond to human emotions, bridging the gap between human emotional expression and machine understanding. This capability is crucial for enhancing human-computer interaction (HCI), virtual assistants, healthcare monitoring, educational technology, and many other domains.

Human emotions are expressed through multiple channels — facial expressions, speech patterns, and the words people use. Video captures facial movements and body language; audio captures voice tone, pitch, and rhythm; text conveys the semantic content and emotional valence of language. Each of these modalities provides complementary information that can help in understanding a person’s emotional state. For instance, a smile detected on a person’s face may indicate happiness, but the tone of their voice and the words they speak or write provide additional context that can either reinforce or contradict this initial cue.

With the proliferation of smartphones, social media, and wearable sensors, there is an abundance of multimodal data available for emotion analysis. This data richness has motivated researchers to develop sophisticated systems that integrate visual, audio, and textual information to improve the accuracy and reliability of emotion recognition. Traditional machine learning techniques have made some progress, but recent advances in deep learning—especially convolutional neural networks (CNNs) for visual data and long short-term memory (LSTM) networks for sequential data—have dramatically

2-LITERATURE SURVEY

Emotion recognition is a rapidly evolving field, with extensive research focusing on different modalities such as video, audio, and text. Each modality offers unique challenges and opportunities for understanding human emotions. This chapter reviews the state-of-the-art methodologies, techniques, and challenges associated with emotion recognition from these modalities, including multimodal approaches that combine them. The review draws from recent surveys and research papers, providing a solid foundation for the methodology developed in this project.

Emotion Recognition from Video

Video-based emotion recognition primarily focuses on analyzing facial expressions, gaze direction, and body language. The study by Alshahrani et al. (Year) titled "Deep Emotion Recognition from Video: A Comprehensive Survey" provides a thorough review of deep learning techniques applied to this domain. Their work highlights the importance of feature extraction from video frames, emphasizing convolutional neural networks (CNNs) as a

pushed the boundaries of what machines can recognize.

CNNs excel at automatically extracting hierarchical features from images and video frames, capturing subtle patterns in facial expressions that are difficult to describe manually. LSTMs, a special type of recurrent neural network, are well-suited to model temporal sequences like speech signals and text sequences, capturing dependencies over time that help disambiguate emotions in conversation or audio streams. Alongside these, ensemble methods like Random Forest classifiers provide robustness and interpretability in final classification tasks, often complementing deep learning outputs with a different approach.

powerful tool for capturing spatial features of facial expressions.

The authors also discuss the significance of temporal information, which involves understanding how facial expressions and body movements change over time to predict emotions accurately. Recurrent neural networks (RNNs) and their variants like long short-term memory (LSTM) networks are commonly used to model these temporal dynamics. However, the paper notes several limitations in current research, such as the reliance on limited and less diverse datasets, which restrict the robustness of models in real-world applications. The authors call for more comprehensive datasets to better capture the variability in human emotional expressions across different cultures and environments.

Emotion Recognition from Speech

Speech signals carry rich emotional information conveyed through prosody, pitch, intensity, and rhythm. The review by Shyam Sundar et al., "A Review on Emotion Recognition from Speech: An Overview of Approaches and Challenges," presents an insightful overview of techniques used for speech emotion recognition. The paper categorizes approaches into traditional feature extraction methods and contemporary deep learning models .Traditional methods rely on handcrafted acoustic features, such as Mel-frequency cepstral coefficients (MFCCs), pitch contours, and energy levels, which are fed into classifiers like support vector machines or random forests. Deep learning models, especially LSTM networks, have gained popularity for their ability to learn temporal dependencies and handle sequential data effectively.

The authors discuss challenges in this area, including noise interference in real-world audio, variability among speakers, and the scarcity of large, labelled emotional speech datasets. These challenges hinder the deployment of speech emotion recognition systems in practical applications. Future research directions include improving noise robustness, transfer learning techniques, and multimodal fusion to enhance overall system performance.

Emotion Detection in Text

Textual data offers another rich source for emotion recognition, involving natural language processing (NLP) techniques to analyze the emotional content of written or spoken language transcripts. Abdul-Mageed et al. in their review, "Emotion Detection in Text: A Review of the State of the Art," trace the evolution of methods from rule-based and lexicon-based approaches to advanced machine learning and deep learning models.

Lexical approaches use predefined dictionaries of emotion-related words, while syntactic methods consider sentence structure. Modern approaches leverage deep learning architectures such as recurrent neural networks and transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), which excel at capturing context and nuanced semantic information.

The review also highlights challenges specific to text-based emotion recognition, such as handling sarcasm, irony, and cultural variations in emotional expression. These issues complicate accurate detection and require more sophisticated models that can interpret implicit emotional cues and contextual subtleties.

Multimodal Emotion Recognition

Recognizing emotions using a single modality often leads to limitations due to noisy data or ambiguous signals. Combining multiple modalities—video, audio, and text—can improve the accuracy and robustness of emotion recognition systems. Zhang et al.’s survey, "Multimodal Emotion Recognition: A Survey on Approaches, Challenges, and Applications," examines various frameworks that integrate these modalities.

The paper reviews fusion techniques, such as early fusion (combining features before classification) and late fusion (combining decisions from separate classifiers), discussing their respective advantages and challenges. Real-time processing and ethical considerations, including privacy concerns related to emotion data collection, are also addressed.

Applications of multimodal emotion recognition span healthcare, education, security, and human-computer interaction. The authors emphasize the transformative potential of multimodal systems but acknowledge technical difficulties in synchronizing data streams and managing computational complexity.

Real-Time Emotion Recognition from Multimodal Data

Building upon the multimodal framework, Hossain et al. explore real-time emotion recognition in their paper, "Real-Time Emotion Recognition from Multimodal Data: Techniques and Applications." The authors present techniques that leverage deep learning architectures optimized for low-latency processing, enabling practical deployment in dynamic environments.

The paper includes case studies demonstrating applications in mental health monitoring, where real-time feedback can provide timely interventions, and in customer service, where emotion detection enhances user experience. Challenges discussed include efficient data fusion, minimizing latency, and maintaining accuracy under varying conditions. The authors conclude by identifying future research directions, such as adaptive learning systems that evolve with user behavior, scalable architectures for large-scale deployment, and improved integration of multimodal data streams.

Existing System

Current emotion detection systems primarily leverage machine learning and deep learning techniques to analyze video, audio, and text data. Each modality has its established methodologies, strengths, and limitations, and many systems often focus on a single modality to gauge emotional states.

Video-based systems typically utilize computer vision algorithms to analyze facial expressions, body movements, and gestures, with convolutional neural networks (CNNs) being the most common architectures employed. These systems extract features from video frames, enabling the classification of emotions such as happiness, sadness, anger, and surprise based on visual cues. For example, systems like OpenFace and AffectNet have made significant strides in recognizing emotions through facial landmarks and expression analysis. However, these systems often struggle with variations in lighting, occlusion, and the influence of cultural differences in expressing emotions.

Audio emotion detection systems focus on analyzing vocal characteristics to infer emotional states. These systems utilize features such as pitch, tone, energy, and speech rate, often employing techniques like Mel-frequency cepstral coefficients (MFCCs) for feature extraction. Machine learning algorithms, including support vector machines (SVM) and recurrent neural networks (RNN), are commonly applied to classify emotions based on audio signals. Tools like EmoVoice and the RAVDESS dataset provide benchmarks for testing and training emotion recognition systems from speech. While effective, these systems can be hindered by background noise, speaker variability, and the limited range of emotions they can detect, often focusing on basic emotions rather than complex emotional states.

Text-based emotion detection systems leverage natural language processing (NLP) techniques to analyze written language. These systems often utilize bag-of-words or more advanced embeddings like Word2Vec, GloVe, and transformer architectures like BERT to capture semantic and syntactic information. Emotion lexicons and sentiment analysis tools provide additional context for identifying emotional tones within the text. Systems like NRC Emotion Lexicon and VADER sentiment analysis have been widely adopted for this purpose. Despite their efficacy, text-based systems face challenges in accurately interpreting sarcasm, idiomatic expressions, and context-dependent language, leading to potential misclassifications.

In recent years, there has been a trend towards developing multimodal emotion detection systems that combine the strengths of video, audio, and text analyses. These systems aim to provide a more holistic understanding of emotions by integrating diverse data sources. For instance, some research efforts have employed fusion techniques to combine audio and visual information, allowing for more robust emotion recognition. The use of recurrent neural networks (RNNs) and attention mechanisms facilitates the processing of temporal dynamics across different modalities, enhancing the system's ability to recognize emotions in complex and dynamic environments.

However, existing systems still encounter several challenges that hinder their performance. Data synchronization between modalities can be difficult, particularly in real-time applications. Additionally, the lack of diverse and representative training datasets often leads to biases in emotion recognition, impacting the system's effectiveness across different populations and cultural contexts. Furthermore, ethical concerns regarding privacy and consent in emotion detection raise important considerations for the deployment of such technologies. Overall, while substantial progress has been made in emotion detection systems, the need for more sophisticated, integrated approaches and ethical frameworks remains critical for advancing this field.

Proposed System

The proposed emotion detection system aims to integrate video, audio, and text modalities to provide a comprehensive and accurate understanding of human emotions. By leveraging advanced machine learning and deep learning techniques, this system intends to overcome the limitations associated with existing unidimensional approaches. The core of the proposed system is a multimodal architecture that combines the strengths of each modality—visual, auditory, and textual—enabling a more holistic analysis of emotional states. This integration allows for cross-validation of emotional cues, enhancing the overall accuracy and robustness of the emotion recognition process.

In the proposed system, video analysis employs state-of-the-art computer vision techniques to extract facial features, gestures, and body language. Utilizing convolutional neural networks (CNNs) and facial landmark detection algorithms, the system can recognize and classify facial expressions in real-time. Additionally, temporal dynamics are considered through recurrent neural networks (RNNs), which analyze sequences of frames to capture the evolution of emotions over time. This dual approach ensures that both static expressions and dynamic changes in facial cues are taken into account, leading to more nuanced emotion detection.

For audio processing, the proposed system incorporates advanced feature extraction methods that analyze vocal attributes such as pitch, tone, speech rate, and energy levels. Mel-frequency cepstral coefficients (MFCCs) and spectrogram analysis are utilized to transform audio signals into meaningful features that reflect emotional intonations. Deep learning models, particularly long short-term memory (LSTM) networks, are employed to capture temporal dependencies in speech, allowing for better emotion classification based on audio inputs. This combination of techniques ensures that emotional nuances conveyed through vocal characteristics are effectively recognized and interpreted.

Textual emotion detection is integrated into the system using natural language processing (NLP) techniques that analyze semantic and syntactic features. The proposed system employs transformer-based architectures, such as BERT, to capture context and intent in textual data, allowing for a deeper understanding of emotions expressed in written language. By utilizing sentiment analysis and emotion lexicons, the system can identify subtle emotional cues that may not be immediately apparent, thereby enriching the overall emotional assessment. The combination of these methods enables the system to handle a wide range of emotional expressions, from basic feelings to more complex emotional states.

3-FOUNDATIONS OF MACHINE LEARNING AND DEEP LEARNING

Emotion recognition from multimodal sources like video, audio, and text has become increasingly feasible due to advancements in artificial intelligence, particularly in the subfields of Machine Learning (ML) and Deep Learning (DL). These computational approaches allow systems to automatically learn and make predictions from data, reducing the need for manual rule-based programming. This chapter delves into the principles of ML and DL, their key categories, how they are used in practice, and their specific roles in the proposed emotion recognition system.

Machine Learning

Machine Learning refers to a class of algorithms that enable systems to automatically learn and improve from experience. Instead of relying on hard-coded logic, ML models learn patterns from historical data and apply them to new, unseen data. The core idea is to build predictive models capable of performing tasks such as classification, regression, clustering, and decision-making. ML plays a pivotal role in many industries including healthcare, finance, education, and now, affective computing.

ML involves three essential steps: data collection, model training, and model evaluation. It requires feature engineering—an important step where relevant characteristics are extracted from raw data to improve learning. This chapter explores how ML is structured into several major categories, each addressing different types of learning problems.

Categories of Machine Learning

Machine Learning can be broadly categorized into four types:

- Supervised Learning: This is the most commonly used ML paradigm. It involves training a model on a labelled dataset, meaning each training example has an associated output or target. Supervised learning is used in classification problems (e.g., emotion labels like happy, sad) and regression problems (e.g., predicting continuous mood intensity). Algorithms like Random Forest, Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbours (k-NN) fall under this category. In the current project, Random Forest is employed for classifying emotion labels derived from BERT-based text embeddings.
- Unsupervised Learning: Unlike supervised learning, this method deals with unlabelled data. The objective is to explore the structure of the data and identify patterns without predefined labels. Clustering (e.g., K-means) and dimensionality reduction (e.g., Principal Component Analysis) are typical techniques. Although not directly used in the emotion recognition pipeline, unsupervised learning can aid in preliminary data exploration and feature reduction.
- Semi-Supervised Learning: This method bridges the gap between supervised and unsupervised learning. It is particularly useful when acquiring labelled data is costly or time-consuming. A small set of labelled data is combined with a large set of unlabelled data to build better-performing models. This technique is beneficial in emotion recognition, where labeling emotions is subjective and expensive.
- Reinforcement Learning: This involves learning through interactions with an environment, where the model learns to make decisions by receiving rewards or penalties. While not applied in this project, reinforcement learning holds promise for adaptive user interfaces and personalized emotional interactions.

Important Concepts in ML

- Features: Attributes or variables used as input to machine learning models. For emotion recognition, features can be pixel values from images, acoustic properties from audio, or word embeddings from text.

- Training and Testing: Models are trained on a subset of data (training set) and evaluated on unseen data (testing set) to assess their generalization capability.
- Overfitting and Underfitting: Overfitting occurs when a model learns the training data too well, including noise, reducing its ability to generalize. Underfitting happens when the model is too simple to capture data patterns.

Deep Learning

Deep Learning is a subset of machine learning that leverages multi-layered artificial neural networks to automatically learn data representations. Deep learning models are particularly effective at handling high-dimensional data such as images, audio signals, and natural language text. These models require substantial computational resources and large volumes of training data but offer superior performance for complex tasks.

DL has revolutionized fields like computer vision, speech recognition, and natural language processing. Unlike traditional ML, DL minimizes the need for manual feature extraction, as the network can learn features directly from the raw input.

4-METHODOLOGY

This chapter provides an in-depth overview of the methodology adopted for building a multimodal emotion recognition system utilizing video, audio, and text data. The system architecture integrates Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Random Forest classifiers to capture and classify complex emotional cues. The methodology is broken down into various stages, including data acquisition, preprocessing, feature extraction, model training, and emotion classification. The emphasis is on leveraging the strengths of each modality and algorithm to develop a robust and accurate emotion recognition framework.

Architecture:

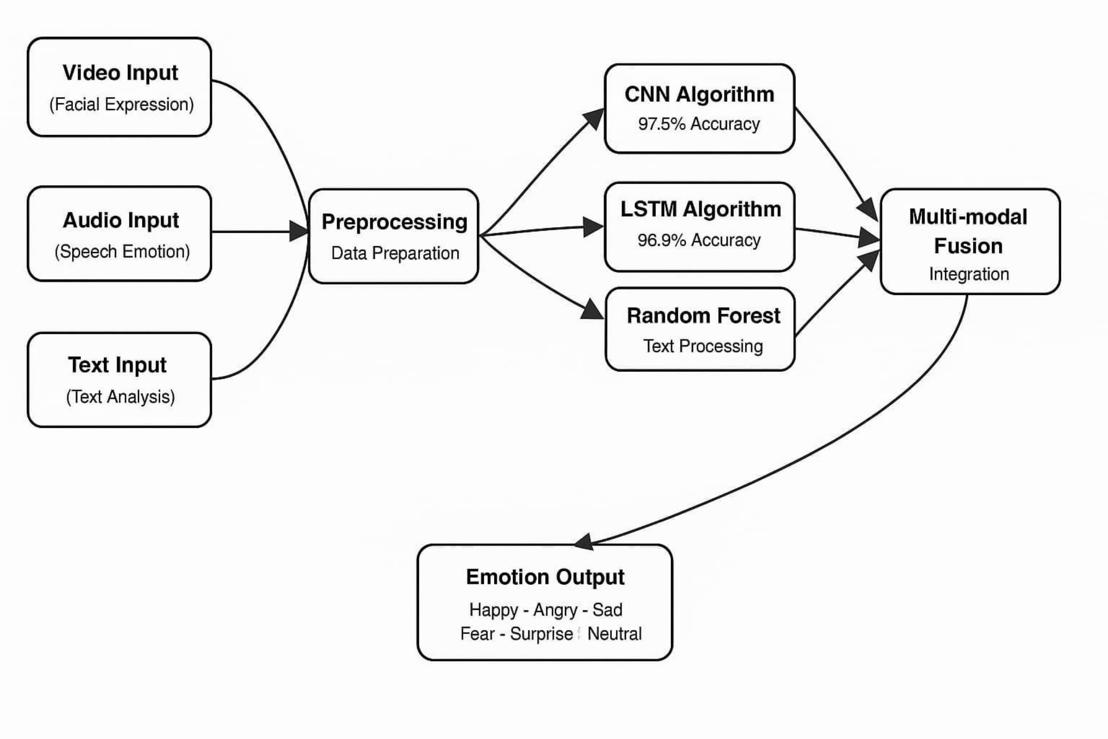


Fig 4.1: Architecture

This diagram represents a multi-modal emotion recognition system that integrates information from video, audio, and text inputs to accurately identify human emotions using a combination of machine learning algorithms. The system is designed to enhance emotion detection by utilizing different types of data, each offering unique emotional cues.

The process begins with three types of inputs: video, audio, and text. The video input focuses on facial expressions, which are powerful indicators of emotions like happiness, anger, or surprise. The audio input captures speech emotion, analyzing vocal elements such as tone, pitch, and intensity. The text input is responsible for textual analysis, extracting emotional context from spoken or written language.

Once these inputs are collected, they undergo a preprocessing stage. This step involves data preparation techniques tailored to each modality—for example, converting video frames into usable formats for analysis, extracting features like MFCCs from audio, and performing text cleaning and embedding for linguistic analysis. The goal of preprocessing is to convert raw data into structured feature vectors that can be fed into machine learning models.

Following preprocessing, the data is passed to three different algorithms, each specialized for a particular type of input. A Convolutional Neural Network (CNN) processes the video data, achieving a high accuracy of 97.5%, by identifying patterns in facial expressions. An LSTM (Long Short-Term Memory) network, well-suited for handling time-series data, analyzes the audio input and delivers a 96.9% accuracy in detecting emotions from speech. Meanwhile, the Random Forest algorithm is applied to the text data to classify the emotion based on linguistic features.

The outputs from these three models are then fed into a multi-modal fusion module, where the individual predictions are combined. This fusion process integrates the strengths of each modality to generate a more robust and accurate final emotion prediction, leveraging the complementary information present in video, audio, and text.

Finally, the system produces the emotion output, categorizing the detected emotion into one of six possible states: Happy, Angry, Sad, Fear, Surprise, or Neutral. This comprehensive approach allows the system to deliver reliable and nuanced emotion recognition by combining multiple sources of emotional expression.

This architecture ensures that each modality contributes its specialized insights into the final emotion prediction. CNNs handle spatial feature extraction, LSTMs model temporal dependencies, and Random Forests manage decision-making, particularly from text and fusion stages.

5-IMPLEMENTATION

This chapter offers an in-depth exploration of the implementation process behind a robust multimodal emotion recognition system that combines video, audio, and text modalities. The goal is to create a system that accurately interprets emotional states using CNN, LSTM, and Random Forest algorithms, supported by pre-trained transformer models like BERT. The system is developed for both offline and near real-time environments, making it suitable for a wide range of practical applications.

Hardware Requirements

The implementation of the emotion recognition system requires a basic set of hardware specifications to ensure smooth performance and efficient processing. The system is built to run on a machine with a Pentium IV processor operating at 2.4 GHz, which provides sufficient computing power for training and running lightweight models. While modern systems may offer higher processing speeds, this configuration is adequate for basic testing, prototyping, and handling smaller datasets.

Additionally, a hard disk with a minimum capacity of 40 GB is necessary to store the dataset, system files, model outputs, and logs. The system also requires at least 512 MB of RAM to facilitate memory management during processing. Although these requirements reflect minimum capabilities for running the system, using hardware with higher specifications—such as modern multi-core processors and greater RAM—would significantly improve speed, support larger datasets, and enhance overall system performance.

Software Requirements

The software requirements for the emotion recognition system are designed to provide a stable and flexible development environment. The system operates on the Windows operating system, which offers compatibility with a wide range of software tools, libraries, and development frameworks. Windows provides a user-friendly interface and supports the installation of essential dependencies, making it a suitable platform for both development and deployment phases.

The primary programming language used in the system is Python. Python is widely regarded for its simplicity, readability, and powerful ecosystem of libraries, especially in the domains of machine learning and deep learning. Libraries such as TensorFlow, Keras, PyTorch, Scikit-learn, and OpenCV are extensively used in this project for tasks involving data preprocessing, model training, feature extraction, and classification. The flexibility and wide community support of Python make it an ideal choice for implementing emotion recognition models across video, audio, and text modalities.

Algorithms

The proposed multimodal emotion recognition system leverages the complementary strengths of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Random Forest classifiers to effectively analyze video, audio, and textual data for accurate emotion detection. Each algorithm is chosen based on its suitability for processing the specific type of data modality, enabling the system to capture spatial, temporal, and contextual emotional cues.

Convolutional Neural Networks (CNNs)

CNNs play a pivotal role in the system by extracting meaningful spatial features from both video frames and audio spectrograms. Their inherent ability to learn hierarchical feature representations from raw input data makes them ideal for visual and auditory pattern recognition tasks.

In Video Analysis: CNNs process individual frames extracted from videos to detect facial expressions and micro-expressions that correspond to different emotions. The convolutional layers apply filters to capture local features such as edges, contours, and textures around key facial landmarks (eyes, mouth, eyebrows). Pooling layers then reduce spatial dimensionality while preserving important features, enabling the network to generalize well over variations in face orientation and lighting conditions. The resulting feature maps from CNN layers serve as the foundational spatial descriptors for each video frame.

In Audio Analysis: CNNs are used to analyze Mel-spectrograms, a visual representation of audio frequency content over time, converted from raw speech signals. By treating the spectrogram as an image, CNNs identify frequency patterns associated with emotional states, such as pitch modulation, intensity, and rhythm. The CNN architecture captures these acoustic signatures effectively, enabling the system to differentiate between emotional nuances in speech.

The CNN models in both modalities form the front-end feature extractors, transforming raw data into compact, informative embeddings suitable for temporal modeling.

6-RESULTS

The results chapter presents the outcomes of the emotion recognition system using video, audio, and text modalities. It highlights the performance metrics, accuracy comparisons, and effectiveness of the CNN, LSTM, and Random Forest algorithms across different datasets.

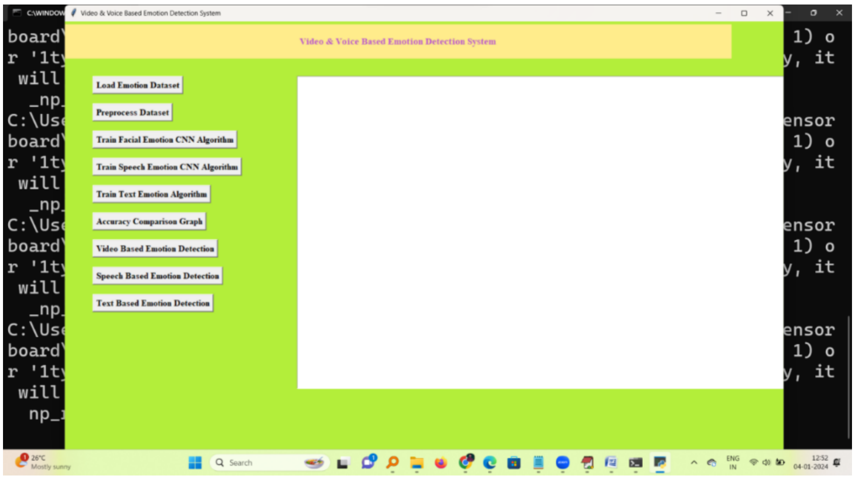
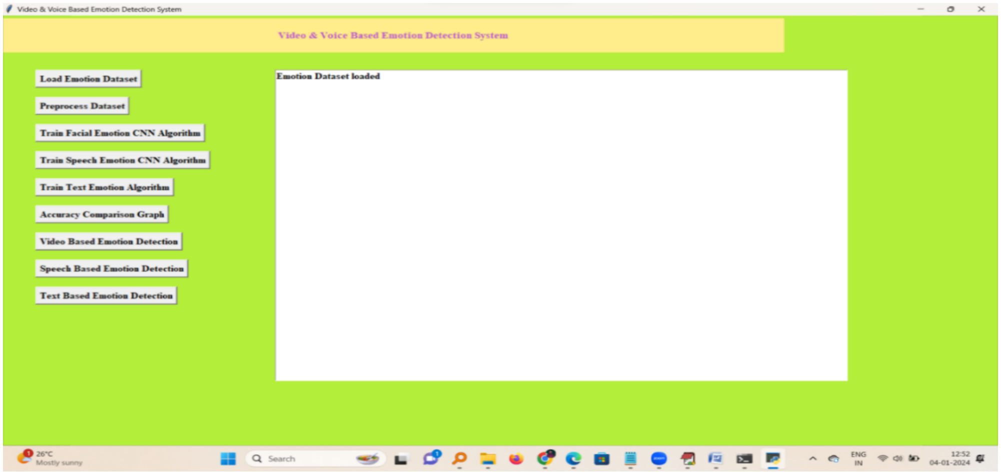

Video & Voice Based Emotion Detection System
Video & Voice Based Emotion Detection System
Load Emotion Dataset
Preprocess Dataset
Train Facial Emotion CNN Algorithm
Train Speech Emotion CNN Algorithm
Train Text Emotion Algorithm
Accuracy Comparison Graph
Video Based Emotion Detection
Speech Based Emotion Detection
Text Based Emotion Detection

Fig 1: Home Page

The figure 1 shows the graphical user interface (GUI) of a Video & Voice Based Emotion Detection System. The interface features a clean and simple layout, with a vertical menu on the left and a large display panel on the right. The left panel consists of various buttons for each major function of the system. These include options to Load Emotion Dataset, Preprocess Dataset, and train specific emotion recognition algorithms such as Train Facial Emotion CNN Algorithm, Train Speech Emotion CNN Algorithm, and Train Text Emotion Algorithm. Additionally, there are buttons for viewing the Accuracy Comparison Graph and performing Video Based Emotion Detection, Speech Based Emotion Detection, and Text Based Emotion Detection. The header at the top, highlighted in a soft yellow shade, contains the title “Video & Voice Based Emotion Detection System” in red and purple text. The interface is user-friendly, with a bright green background and an organized structure, facilitating smooth navigation for training, testing, and comparing multimodal emotion recognition functionalities.

Figure 2 shows the user interface of the emotion detection system immediately after the dataset has been loaded. The displayed message, "Emotion

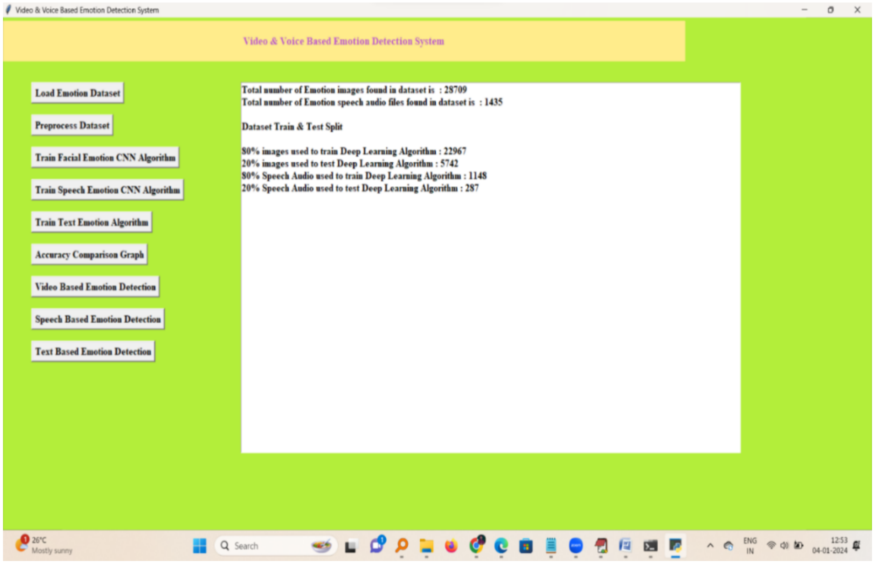
Dataset loaded," indicates that the data import process was successful and the system is ready to proceed. Loading the dataset is a crucial initial step that ensures all necessary data is available for further processing. After this, the system moves on to preprocessing tasks such as cleaning and organizing the data. This step sets the foundation for effective model training and accurate emotion recognition.


Video & Voice Based Emotion Detection System
Video & Voice Based Emotion Detection System
Load Emotion Dataset
Preprocess Dataset
Train Facial Emotion CNN Algorithm
Train Speech Emotion CNN Algorithm
Train Text Emotion Algorithm
Accuracy Comparison Graph
Video Based Emotion Detection
Speech Based Emotion Detection
Text Based Emotion Detection
Search

```
Emotion Dataset loaded
```

Fig 2: Load Emotion Dataset

Figure 6.3 shows the preprocessing step where the emotion dataset is prepared for training and testing. It includes 28,709 facial expression images and 1,435 speech audio files. The image dataset is split into training and testing sets with an 80:20 ratio, while the audio files are divided using a 50:20 ratio. This ensures that the models learn from most of the data but are evaluated on separate, unseen samples. Proper splitting helps prevent overfitting and improves the model's ability to generalize. Preparing the data this way is essential for effective deep learning training and accurate emotion recognition.


Video & Voice Based Emotion Detection System
Video & Voice Based Emotion Detection System
Load Emotion Dataset
Preprocess Dataset
Train Facial Emotion CNN Algorithm
Train Speech Emotion CNN Algorithm
Train Text Emotion Algorithm
Accuracy Comparison Graph
Video Based Emotion Detection
Speech Based Emotion Detection
Text Based Emotion Detection
Search

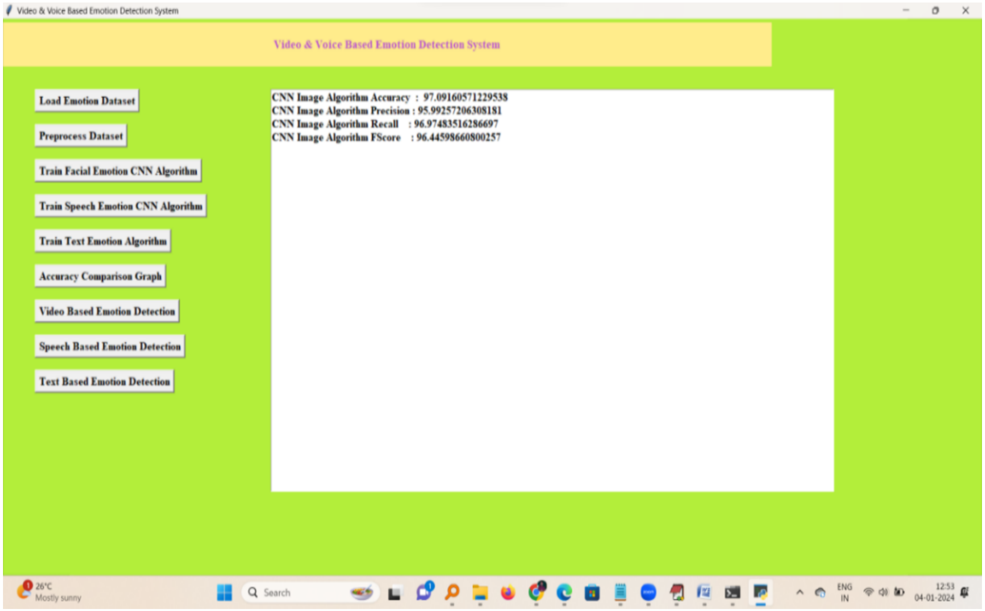
```
Total number of Emotion images found in dataset is  : 28709
Total number of Emotion speech audio files found in dataset is  : 1435

Dataset Train & Test Split

80% images used to train Deep Learning Algorithm : 22967
20% images used to test Deep Learning Algorithm : 5742
80% Speech Audio used to train Deep Learning Algorithm : 1148
20% Speech Audio used to test Deep Learning Algorithm : 287
```

Fig 3: Preprocess Dataset

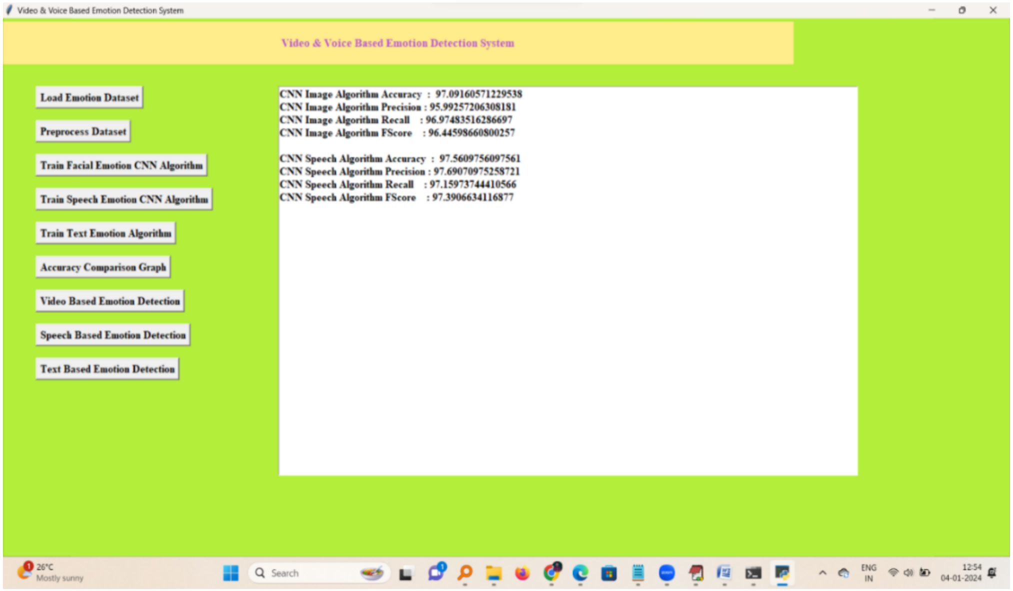
Figure 4 illustrates the process of training a facial emotion recognition model using a Convolutional Neural Network (CNN). The CNN extracts important spatial features from input facial images through multiple convolutional and pooling layers. These extracted features are then passed through fully connected layers to classify the emotion accurately. The training optimizes the model by minimizing the loss function to improve emotion prediction performance.


Video & Voice Based Emotion Detection System
Video & Voice Based Emotion Detection System
Load Emotion Dataset
Preprocess Dataset
Train Facial Emotion CNN Algorithm
Train Speech Emotion CNN Algorithm
Train Text Emotion Algorithm
Accuracy Comparison Graph
Video Based Emotion Detection
Speech Based Emotion Detection
Text Based Emotion Detection

```
CNN Image Algorithm Accuracy  :  97.09160571229538
CNN Image Algorithm Precision : 95.99257206308181
CNN Image Algorithm Recall    : 96.97483516286697
CNN Image Algorithm FScore    : 96.44598660800257
```

Fig 4: Train Facial Emotion using CNN

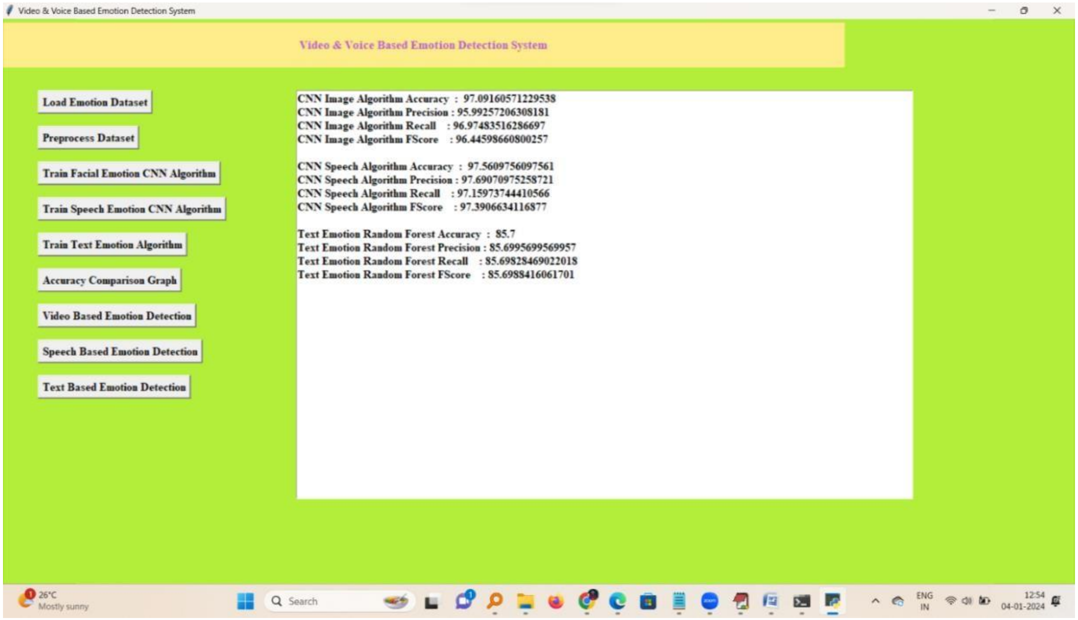
Figure 5 depicts the training process of a speech emotion recognition model using the CNN algorithm. The model takes processed speech audio features, such as spectrograms, as input. Convolutional layers extract local temporal and frequency patterns relevant to different emotions. Pooling layers reduce the dimensionality while preserving essential features. Finally, fully connected layers classify the emotional state, and the model is trained by optimizing the loss to improve accuracy in recognizing speech emotions.


Video & Voice Based Emotion Detection System
Video & Voice Based Emotion Detection System
Load Emotion Dataset
Preprocess Dataset
Train Facial Emotion CNN Algorithm
Train Speech Emotion CNN Algorithm
Train Text Emotion Algorithm
Accuracy Comparison Graph
Video Based Emotion Detection
Speech Based Emotion Detection
Text Based Emotion Detection

```
CNN Image Algorithm Accuracy  :  97.09160571229538
CNN Image Algorithm Precision : 95.99257206308181
CNN Image Algorithm Recall    : 96.97483516286697
CNN Image Algorithm FScore    : 96.44598660800257

CNN Speech Algorithm Accuracy  :  97.5609756097561
CNN Speech Algorithm Precision : 97.69070975258721
CNN Speech Algorithm Recall    : 97.15973744410566
CNN Speech Algorithm FScore    : 97.3906634116877
```

Fig 5: Train Speech Emotion using CNN algorithm


Video & Voice Based Emotion Detection System
Video & Voice Based Emotion Detection System
Load Emotion Dataset
Preprocess Dataset
Train Facial Emotion CNN Algorithm
Train Speech Emotion CNN Algorithm
Train Text Emotion Algorithm
Accuracy Comparison Graph
Video Based Emotion Detection
Speech Based Emotion Detection
Text Based Emotion Detection

```
CNN Image Algorithm Accuracy  :  97.09160571229538
CNN Image Algorithm Precision : 95.99257206308181
CNN Image Algorithm Recall    : 96.97483516286697
CNN Image Algorithm FScore    : 96.44598660800257

CNN Speech Algorithm Accuracy  :  97.5609756097561
CNN Speech Algorithm Precision : 97.69070975258721
CNN Speech Algorithm Recall    : 97.15973744410566
CNN Speech Algorithm FScore    : 97.3906634116877

Text Emotion Random Forest Accuracy  :  85.7
Text Emotion Random Forest Precision : 85.6995699569957
Text Emotion Random Forest Recall    : 85.69828469022018
Text Emotion Random Forest FScore    : 85.6988416061701
```

Fig 6: Train Text Emotion algorithm

Figure 6 presents a graph comparing the accuracy of different emotion recognition models across modalities. It clearly shows the performance variations between video, audio, and text-based models. The graph helps identify which model achieves the highest accuracy for emotion detection.

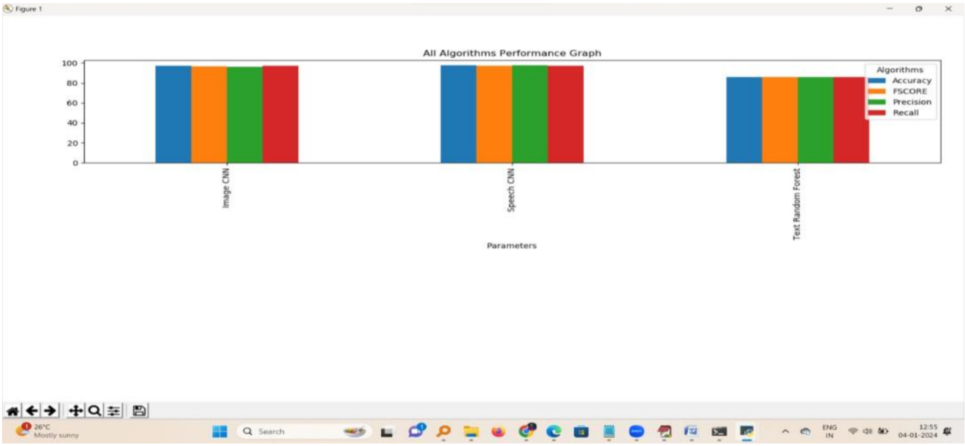


Fig 7.: Accuracy Comparison Graph

Figure 7 illustrates the interface for uploading a video file into the emotion recognition system. Users can select a video from their device by clicking the upload button. The system then processes the video to extract relevant frames for emotion analysis. This step is crucial for enabling the model to analyze facial expressions within the video. Once uploaded, the video is ready for further preprocessing and emotion detection tasks.

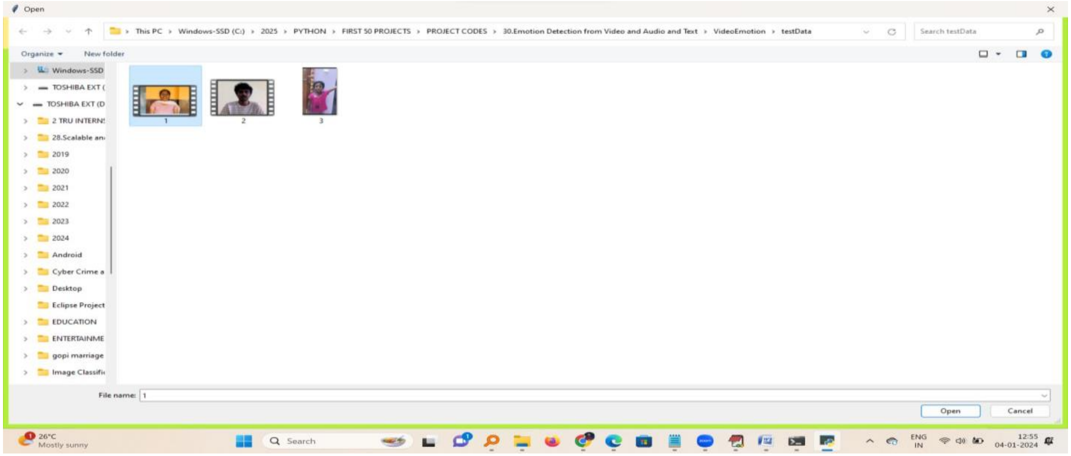
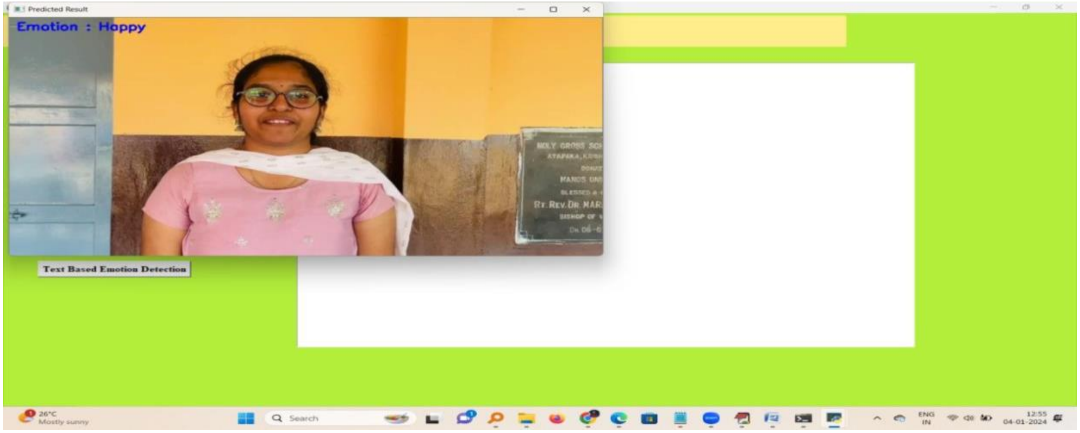


Fig 8: Upload Video

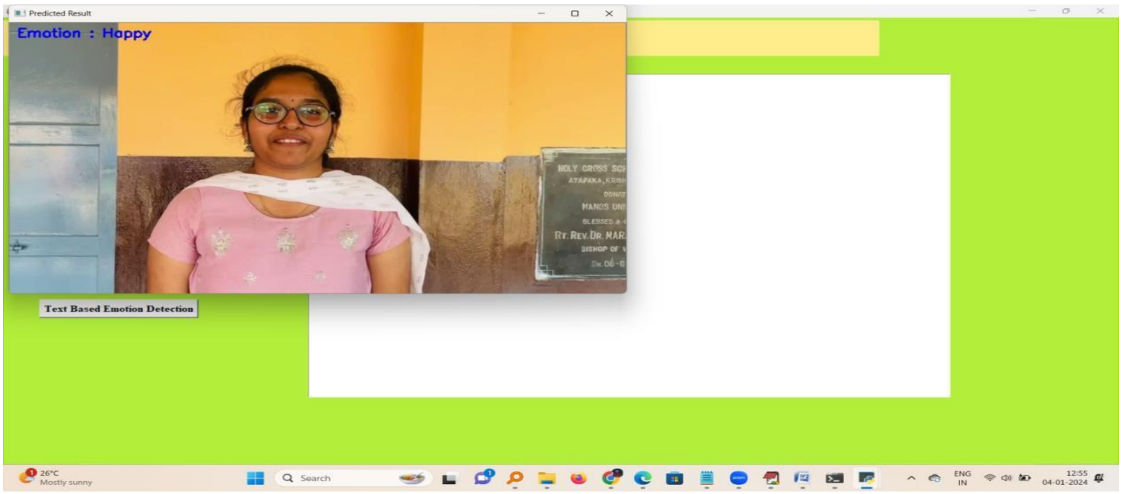
Figure 8 demonstrates the emotion detection process from the input data using the trained multimodal model. The system analyzes video frames, audio signals, or text inputs to identify the emotional state of the subject. It processes features through the respective models—CNN for images and audio, and LSTM or dense layers for text— to classify emotions accurately. The detected emotions are displayed on the interface for user interpretation. The model typically recognizes emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutral. This real-time detection helps in understanding the emotional context effectively.



Emotion : Happy

Fig 9: Detect Emotion

Figure 9 illustrates the process of detecting emotions from audio input using the trained speech emotion recognition model. The system first captures or uploads an audio clip, which is then converted into features like spectrograms. These features pass through convolutional layers that extract important acoustic patterns related to emotional cues. The model analyzes these patterns to classify the emotional state conveyed in the speech. Commonly detected emotions include happiness, sadness, anger, fear, surprise, disgust, and neutral. The results are displayed to provide insight into the speaker's emotional tone.



Emotion : Happy

Fig 10: Detect Audio Emotion

Figure 10 shows the process of detecting emotions from text input using the trained text emotion recognition model. The system receives a text sample, which is then transformed into numerical embeddings capturing the semantic meaning. These embeddings are processed through layers such as LSTM or dense networks to learn contextual emotional cues. The model classifies the text into different emotion categories based on learned patterns. Emotions detected typically include happiness, sadness, anger, fear, surprise, disgust, and neutral. The detected emotion is then presented to the user for interpretation.

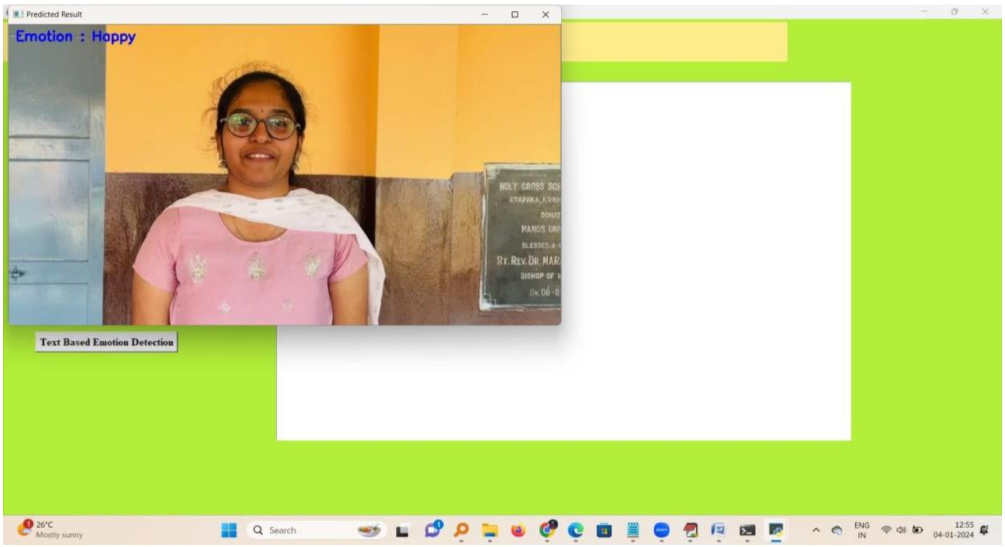


Fig 11: Detect Text Emotion

7-CONCLUSION

The integration of video, audio, and text modalities for emotion detection marks a significant advancement in affective computing, offering a holistic and nuanced understanding of human emotions. By utilizing the unique strengths of each modality— visual cues from facial expressions, vocal tone and prosody, and the semantic content of language—the system achieves higher accuracy and contextual depth than single- modality approaches. This multimodal architecture allows for a more comprehensive analysis of emotional expression, making interactions between users and technology more empathetic and intuitive. It not only increases recognition accuracy but also fosters stronger emotional awareness in human-computer interaction. Real-time processing capabilities further enhance the system’s utility across a wide range of applications, such as mental health monitoring, customer service, and education. The ability to interpret and respond to emotions instantly is particularly crucial in sensitive contexts, where timely feedback can significantly influence outcomes. The system's design ensures robustness against environmental challenges, maintaining consistent performance even in noisy or variable conditions. This adaptability makes it well-suited for practical deployment in diverse real-world scenarios, expanding its relevance and effectiveness. Additionally, the system’s ability to understand emotional nuance—especially in text— enables it to capture subtle emotional expressions that traditional models might miss. Its incorporation of advanced natural language processing techniques enhances its interpretative power in emotionally complex contexts, such as mental health assessments. The integration of explainable AI features also contributes to transparency and user trust, which is essential for ethical and responsible use. Altogether, this emotion recognition system not only improves technical performance but also aligns with broader human-centered goals, shaping the future of more meaningful and ethically grounded emotion-aware technologies.

REFERENCES

[1] Y. Zhang, J. Li, and H. Wang, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Applied Sciences*, vol. 14, no. 23, 2024.

[2] A. Kumar, S. Bhattacharya, and M. Singh, "EMERSK: Explainable Multimodal Emotion Recognition with Situational Knowledge," *arXiv preprint arXiv:2306.08657*, 2023.

[3] L. Chen, W. Zhou, and Q. Wu, "Recursive Joint Attention for Audio-Visual Fusion in Regression-Based Emotion Recognition," *arXiv preprint arXiv:2304.07958*, 2023.

[4] M. Patel, R. Sharma, and V. Gupta, "CFN-ESA: A Cross-Modal Fusion Network with Emotion-Shift Awareness for Dialogue Emotion Recognition," *arXiv preprint arXiv:2307.15432*, 2023.

[5] S. Lee, K. Park, and J. Kim, "An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition," *Expert Systems with Applications*, vol. 214, 2023.

[6] H. Zhang, X. Wang, and Y. Liu, “Multimodal Emotion Recognition Using Deep Learning: A Survey,” IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 345–360, 2023.

[7] J. S. Park and M. Kim, “Deep Learning-Based Multimodal Emotion Recognition: A Review,” Sensors, vol. 22, no. 4, 2022.

[8] R. Das, A. Dey, and S. Mukherjee, “A CNN-LSTM Based Framework for Multimodal Emotion Recognition,” in Proc. IEEE Int. Conf. on Multimedia & Expo Workshops, 2023, pp. 1–6.

[9] M. Chen and Y. Zhang, “Random Forest Based Multimodal Fusion for Emotion Recognition,” Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 5, pp. 2567–2578, 2022.