Multimodal Emotion Recognition

by - Madhav Deshatwad (142402008)

- Shrikant Budde (142402010)

under the guidance of - Dr. Sahely bhadra

date-16 april 2025

IIT PALAKKAD

How many emotions are there?

01

From 20 to 34 k+

There's no definitive number of emotions. While some theories propose a limited set of basic emotions, others suggest a much broader range. The number of emotions depends on how they are defined and categorized.

02

Basic Emotions

Theories like those by Paul Ekman propose a core set of emotions that are universally recognized and expressed, including happiness, sadness, anger, fear, surprise, and disgust.

03

complex Emotions

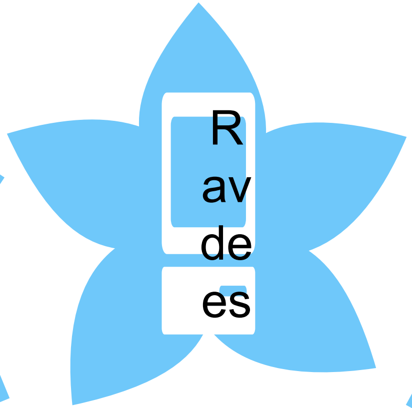
beyond basic emotions ,there are complex emotions that combine multiple feelings,such as jealousy,guilt,pride ,and gratitude.these emotions are often influenced by cultural and personal experiences

Intro To Dataset

Ravdees Video Speech

Features

Modality and Vocal channel are constant as we used speech video dataset.



Emotion

Neutral, calm,

happy, sad,

angry, fearful,

disgust, surprised

Actor

01 to 24.

Odd numbered actors are male,

even numbered actors are female

Repetition

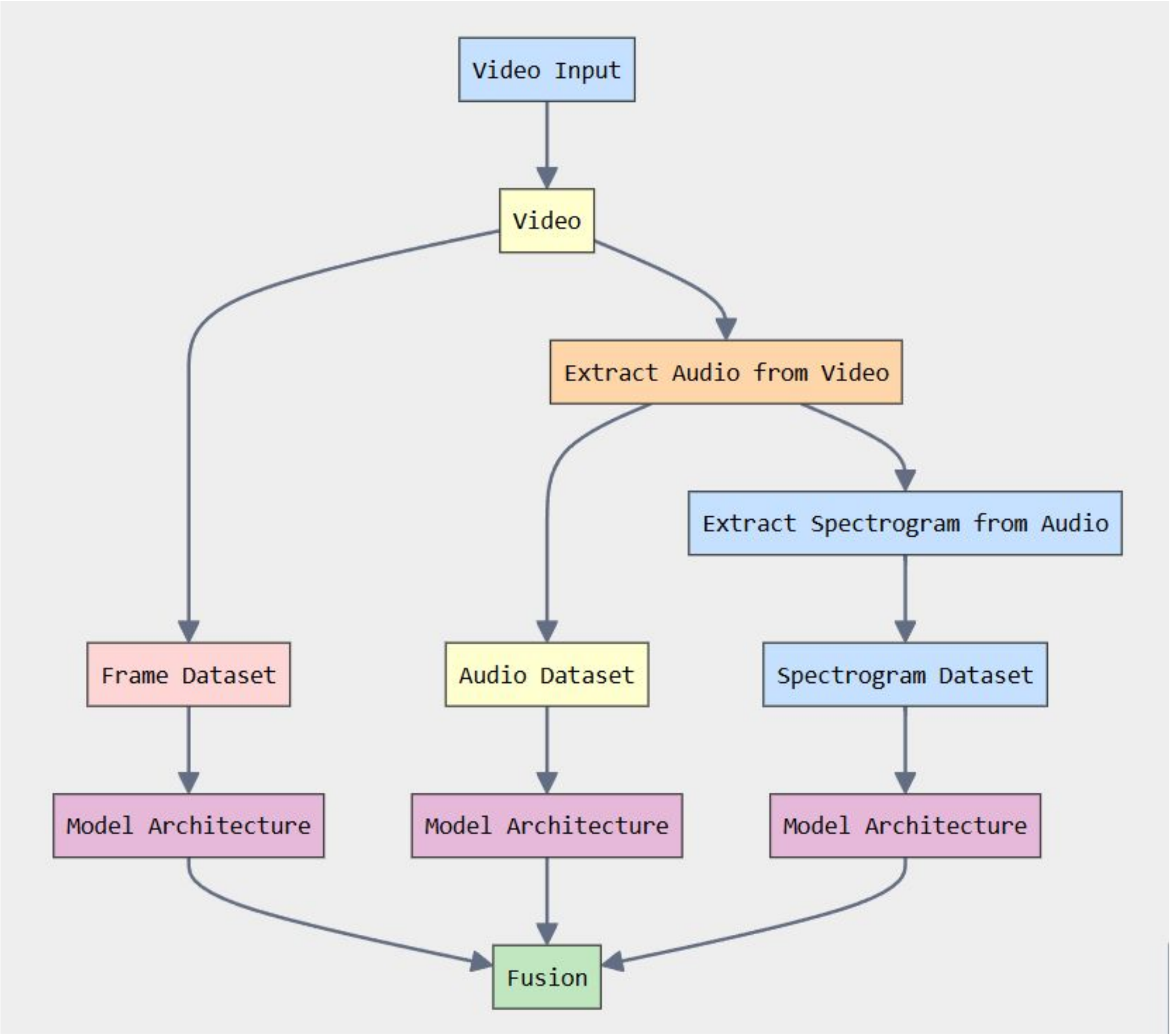
01 = 1st repetition,

02 = 2nd repetition

Statement

- "Kids are talking by the door",

- "Dogs are sitting by the door"


Video Input
Video
Extract Audio from Video
Extract Spectrogram from Audio
Frame Dataset
Audio Dataset
Spectrogram Dataset
Model Architecture
Model Architecture
Model Architecture
Fusion

Multimodal Data Handling & Preparation Stages

1 **Ravdess data**

2 **images and audio data extraction from video**

3 **EDA**

4 **Train,test,vali date**

840,288,312

Video Frame-Based Emotion Classification using ResNet50

Data Acquisition

Collected video samples from RAVDESS dataset with diverse emotional expressions.

Preprocessing

Extracted video frames; applied face detection and normalization techniques.

Model Architecture

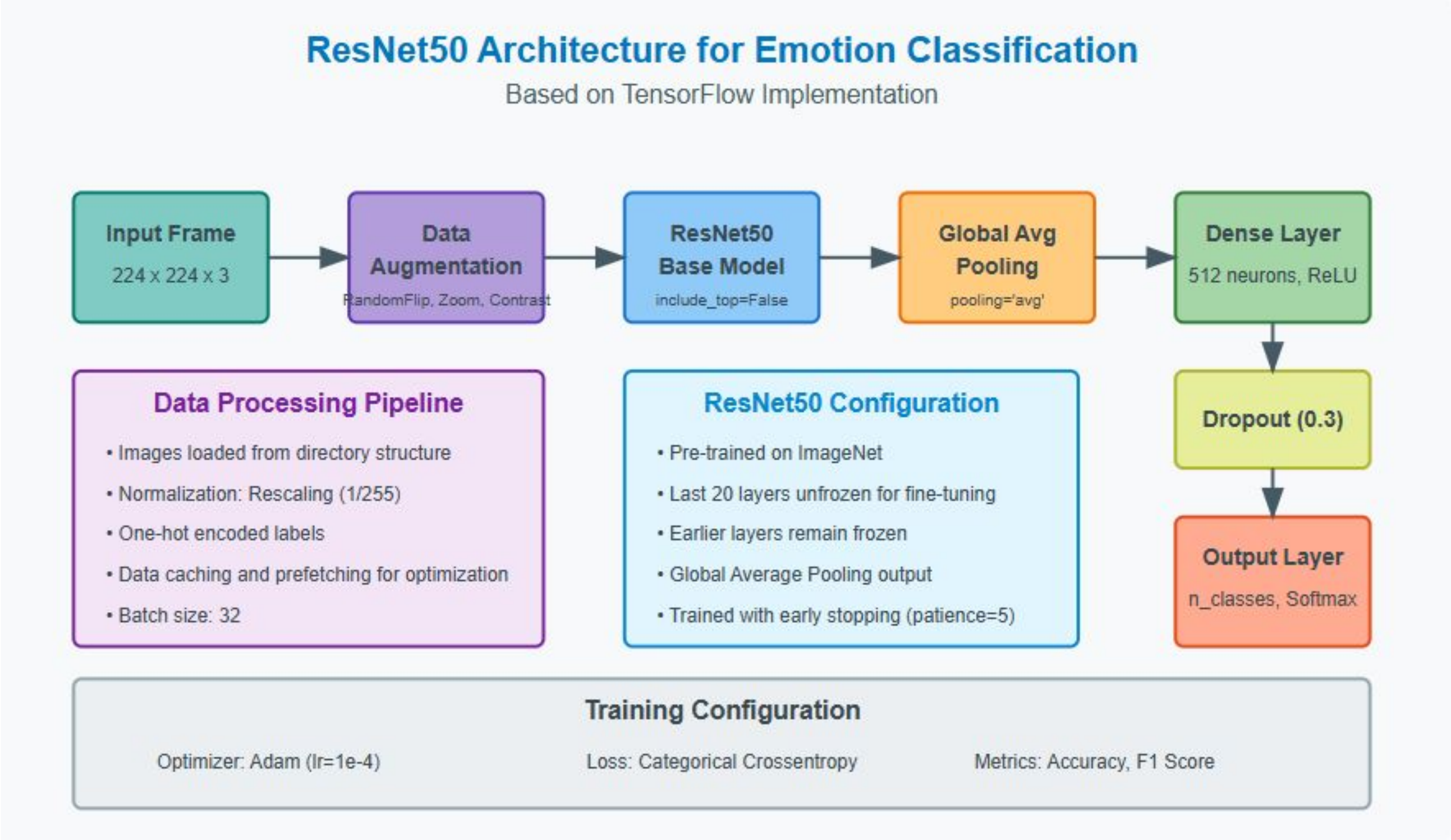
Used pretrained ResNet 50 model, fine-tuned for emotion classification.

Training

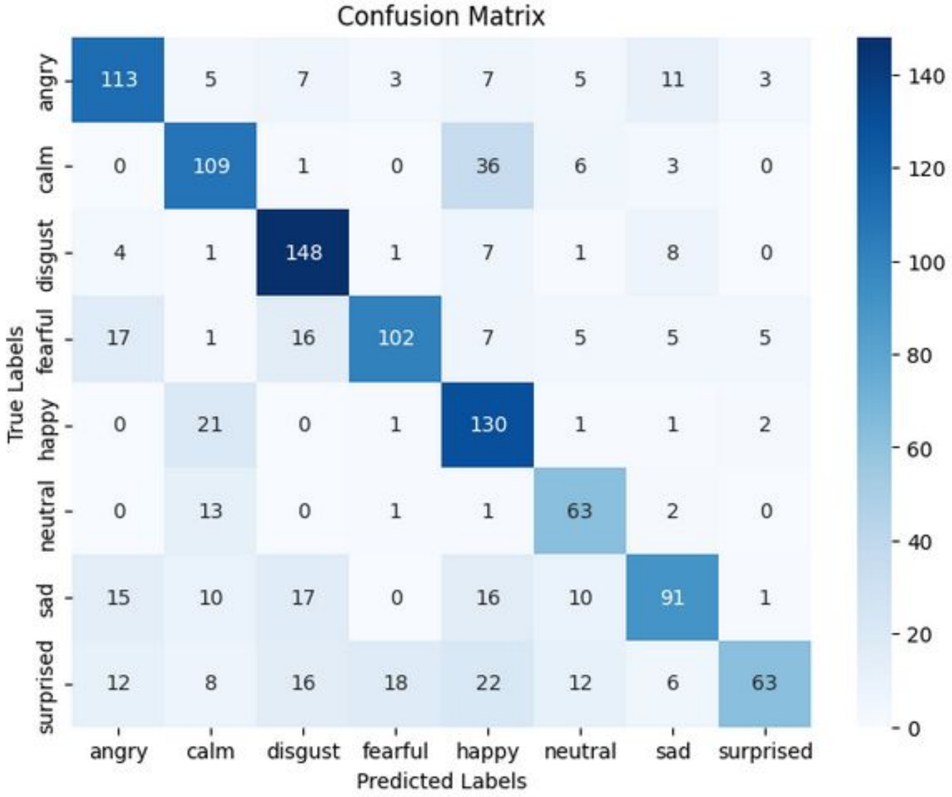
Model trained using cross-entropy loss and Adam optimizer.

Evaluation

Accuracy, confusion matrix, and F1-score used for performance evaluation.



ResNet50 Architecture for Emotion Classification
Based on TensorFlow Implementation
Input Frame
224 x 224 x 3
Data Augmentation
RandomFlip, Zoom, Contrast
ResNet50 Base Model
include_top=False
Global Avg Pooling
pooling='avg'
Dense Layer
512 neurons, ReLU
Dropout (0.3)
Output Layer
n_classes, Softmax
Data Processing Pipeline
• Images loaded from directory structure
• Normalization: Rescaling (1/255)
• One-hot encoded labels
• Data caching and prefetching for optimization
• Batch size: 32
ResNet50 Configuration
• Pre-trained on ImageNet
• Last 20 layers unfrozen for fine-tuning
• Earlier layers remain frozen
• Global Average Pooling output
• Trained with early stopping (patience=5)
Training Configuration
Optimizer: Adam (lr=1e-4)
Loss: Categorical Crossentropy
Metrics: Accuracy, F1 Score



Classification Report:

	precision	recall	f1-score	support
angry	0.70	0.73	0.72	154
calm	0.65	0.70	0.67	155
disgust	0.72	0.87	0.79	170
fearful	0.81	0.65	0.72	158
happy	0.58	0.83	0.68	156
neutral	0.61	0.79	0.69	80
sad	0.72	0.57	0.63	160
surprised	0.85	0.40	0.55	157
accuracy			0.69	1190
macro avg	0.70	0.69	0.68	1190
weighted avg	0.71	0.69	0.68	1190

Audio-Based Emotion Classification using CNN, LSTM, BiLSTM & Attention

Audio Extraction

Extracted audio tracks from RAVDESS video files using standard libraries.

Preprocessing

Extracted video frames; applied face detection and normalization techniques.

Model Architecture

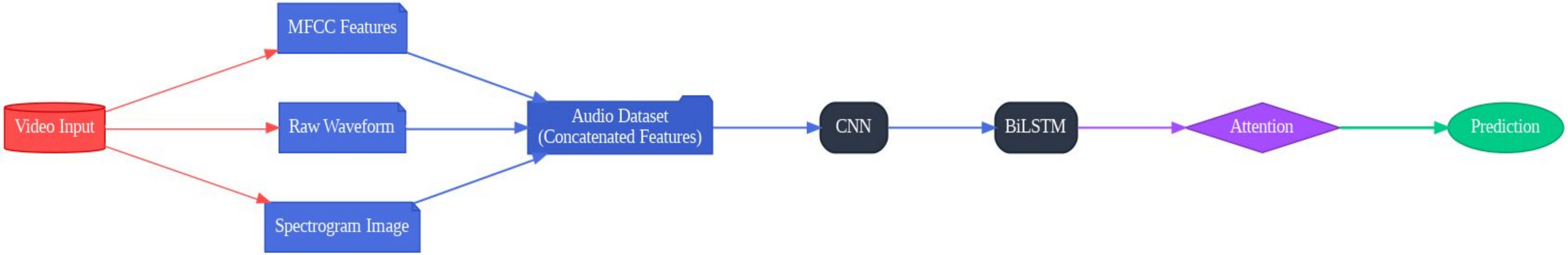
Implemented CNN, LSTM, BiLSTM, and Attention-based models for emotion classification

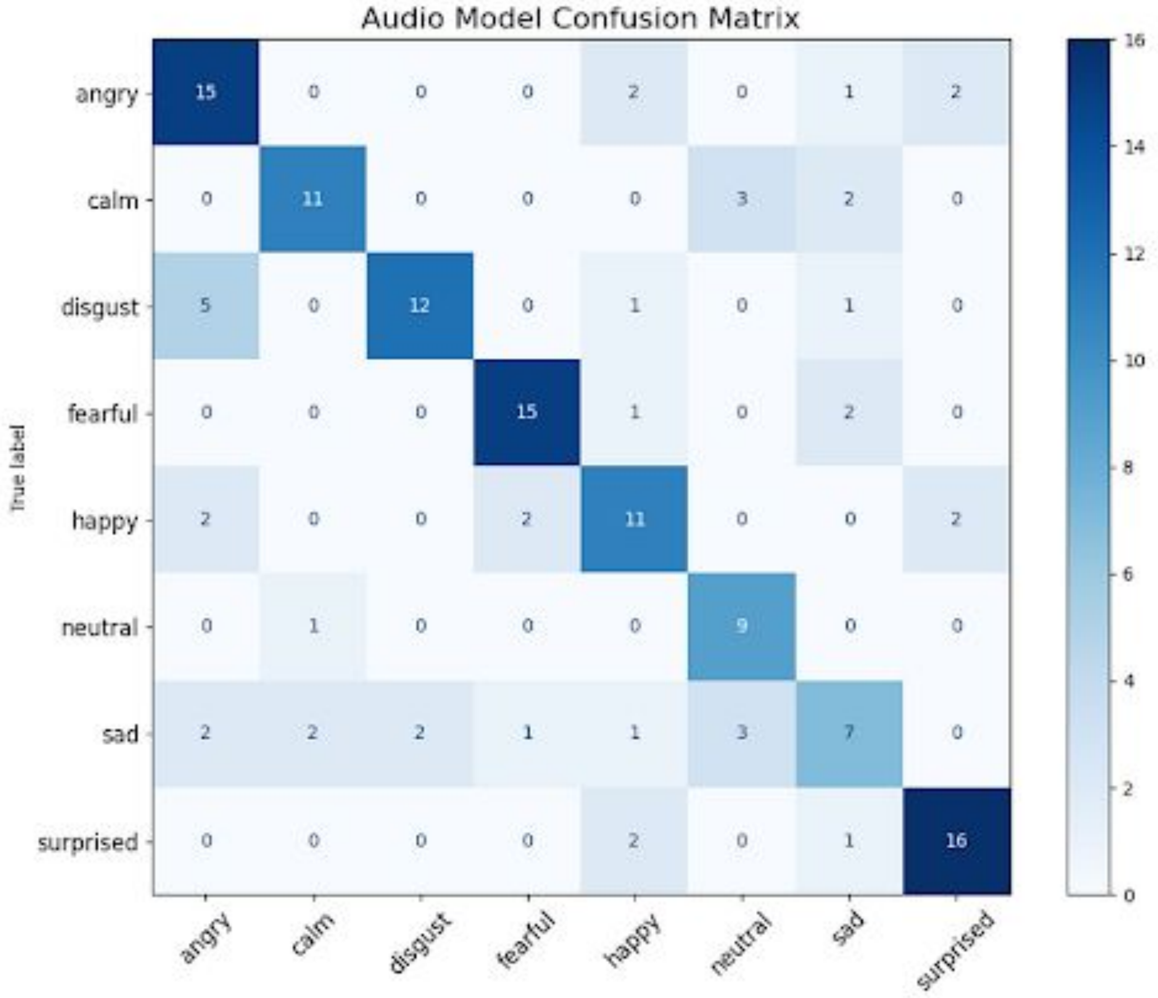
Training

Trained using categorical cross-entropy with data augmentation techniques.

Evaluation

Evaluated using accuracy, precision, recall, F1-score, and ROC curves


Video Input
MFCC Features
Raw Waveform
Spectrogram Image
Audio Dataset
(Concatenated Features)
CNN
BiLSTM
Attention
Prediction



Audio Classification Report:

	precision	recall	f1-score	support
angry	0.62	0.75	0.68	20
calm	0.79	0.69	0.73	16
disgust	0.86	0.63	0.73	19
fearful	0.83	0.83	0.83	18
happy	0.61	0.65	0.63	17
neutral	0.60	0.90	0.72	10
sad	0.50	0.39	0.44	18
surprised	0.80	0.84	0.82	19
accuracy			0.70	137
macro avg	0.70	0.71	0.70	137
weighted avg	0.71	0.70	0.70	137

Audio Spectrogram using Resnet18

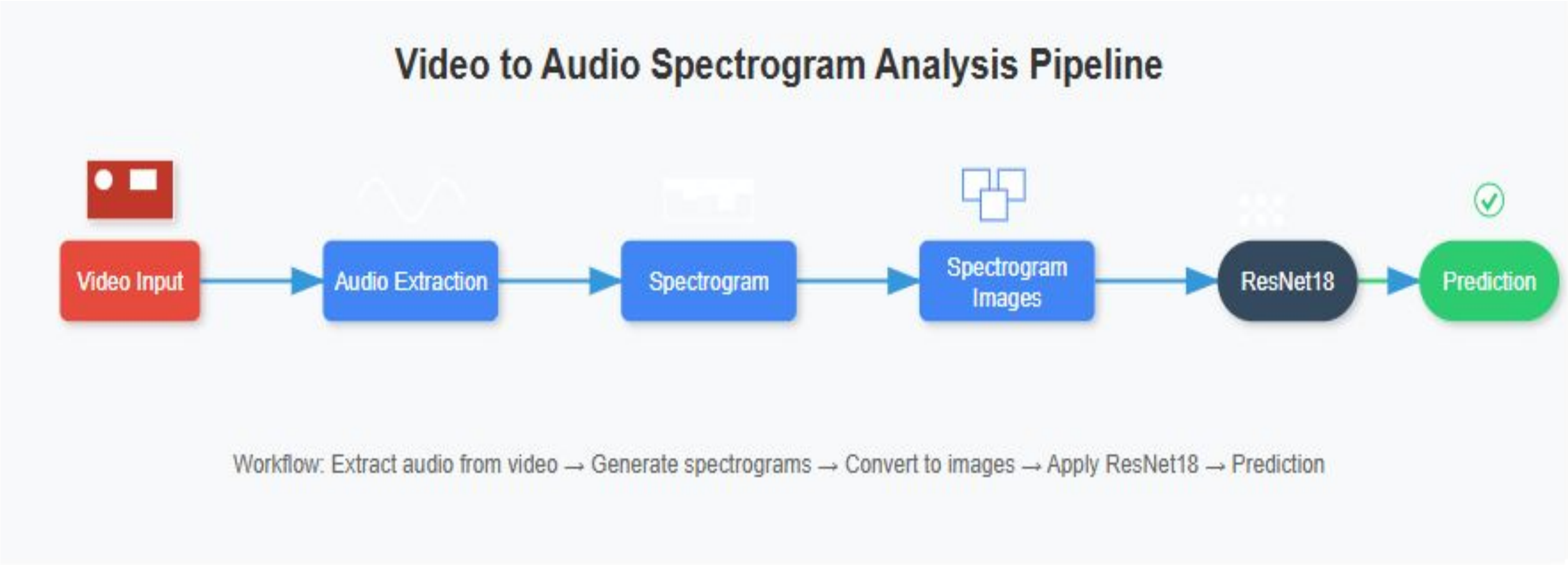
Input: Video files (MP4, AVI, etc.)

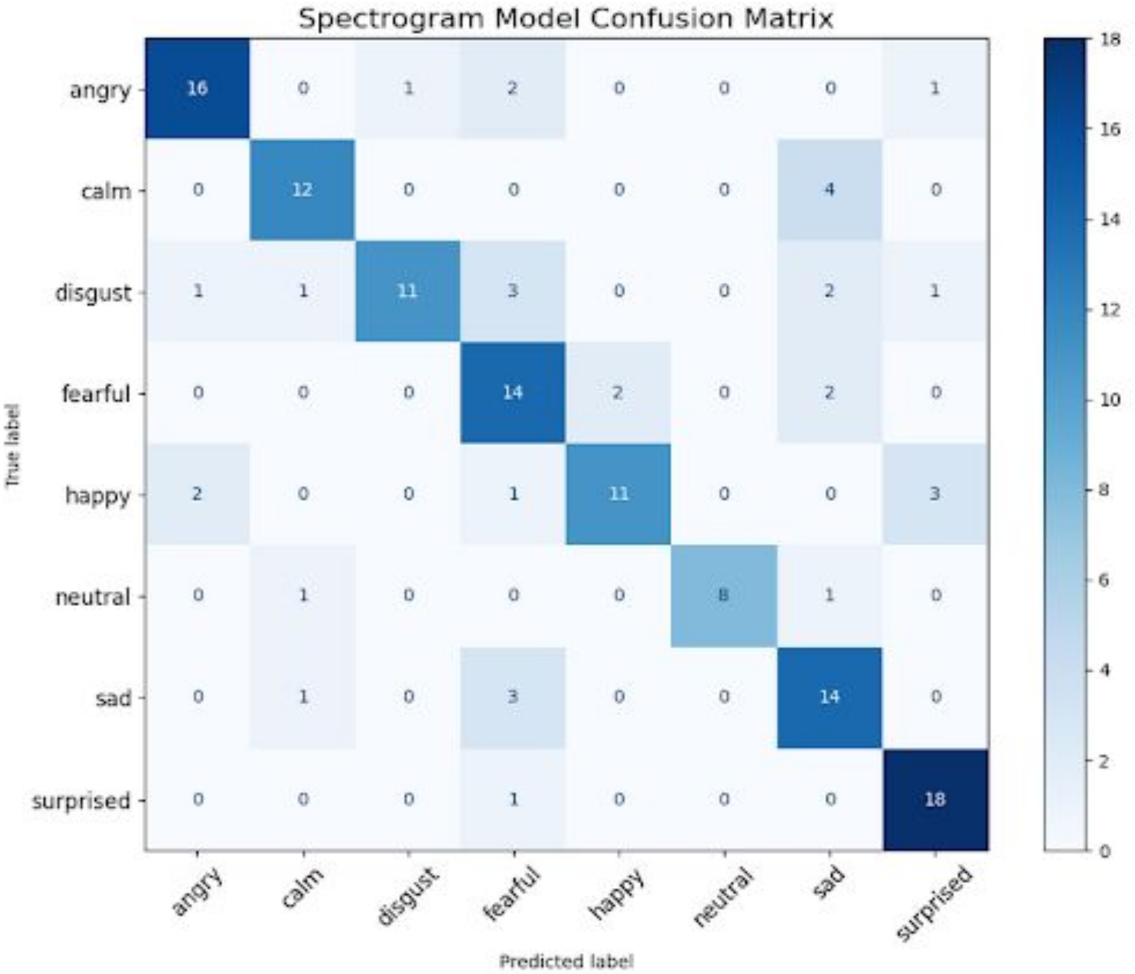
Audio Extraction: Librosa/PyAV for audio track isolation

Spectrograms: Time-frequency representations (Mel-spectrograms shown as 2D images)

ResNet18: Pre-trained CNN adapted for spectrogram analysis

Prediction: Class probabilities


Video to Audio Spectrogram Analysis Pipeline
Video Input
Audio Extraction
Spectrogram
Spectrogram Images
ResNet18
Prediction
Workflow: Extract audio from video → Generate spectrograms → Convert to images → Apply ResNet18 → Prediction



Spectrogram Classification Report:

	precision	recall	f1-score	support
angry	0.84	0.80	0.82	20
calm	0.80	0.75	0.77	16
disgust	0.92	0.58	0.71	19
fearful	0.58	0.78	0.67	18
happy	0.85	0.65	0.73	17
neutral	1.00	0.80	0.89	10
sad	0.61	0.78	0.68	18
surprised	0.78	0.95	0.86	19
accuracy			0.76	137
macro avg	0.80	0.76	0.77	137
weighted avg	0.79	0.76	0.76	137

Text-Based Emotion Recognition

Dataset Splitting

Dataset split into 70% train, 20% validation, 10% test.

Emotion Balancing

Maintained class balance across emotion categories during split.

Text Preprocessing

Applied tokenization, padding, and sequence conversion.

Classification Model

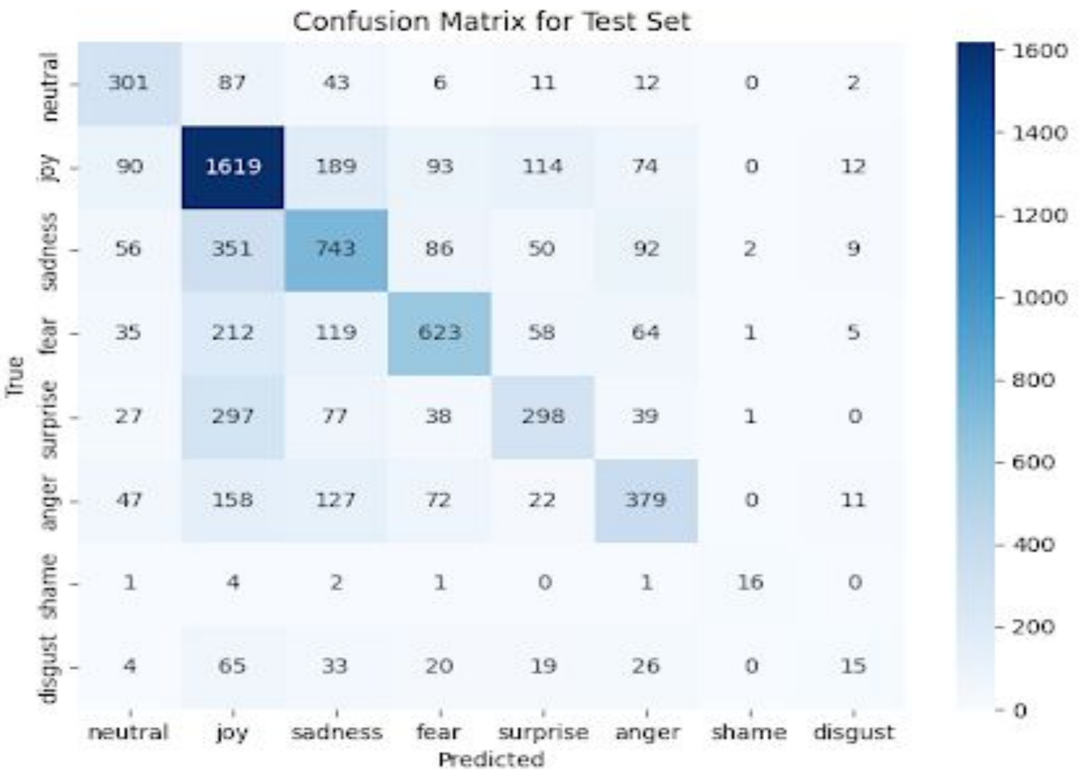
Built models with logistic regression,decision tree,RF,naive bayes

Tuning

Cross Validation and Hyperparameter tuning

Performance Metrics

Evaluated with accuracy, precision, recall, F1-score.



	precision	recall	f1-score	support
neutral	0.54	0.65	0.59	462
joy	0.58	0.74	0.65	2191
sadness	0.56	0.53	0.55	1389
fear	0.66	0.56	0.61	1117
surprise	0.52	0.38	0.44	777
anger	0.55	0.46	0.50	816
shame	0.80	0.64	0.71	25
disgust	0.28	0.08	0.13	182
accuracy			0.57	6959
macro avg	0.56	0.51	0.52	6959
weighted avg	0.57	0.57	0.56	6959

Naive Bayes Model - Training Metrics:
Training Accuracy: 0.5794
Training F1 Score: 0.5717
Training Precision: 0.5817
Training Recall: 0.5794
Naive Bayes Model - Validation Metrics:
Validation Accuracy: 0.5414
Validation F1 Score: 0.5319
Validation Precision: 0.5389
Validation Recall: 0.5414
Naive Bayes Model - Test Metrics:
Test Accuracy: [illegible]
Test F1 Score: [illegible]
Test Precision: [illegible]
Test Recall: [illegible]

Decision Tree Model - Training Metrics:
Training Accuracy: 0.9799
Training F1 Score: 0.9806
Training Precision: 0.9823
Training Recall: 0.9799
Decision Tree Model - Validation Metrics:
Validation Accuracy: 0.5122
Validation F1 Score: 0.5096
Validation Precision: 0.5149
Validation Recall: 0.5122
Decision Tree Model - Test Metrics:
Test Accuracy: [illegible]
Test F1 Score: [illegible]
Test Precision: [illegible]
Test Recall: [illegible]

Random Forest Model - Training Metrics:
Training Accuracy: 0.9798
Training F1 Score: 0.9803
Training Precision: 0.9816
Training Recall: 0.9798
Random Forest Model - Validation Metrics:
Validation Accuracy: 0.5686
Validation F1 Score: 0.5582
Validation Precision: 0.5673
Validation Recall: 0.5686
Random Forest Model - Test Metrics:
Test Accuracy: [illegible]
Test F1 Score: [illegible]
Test Precision: [illegible]
Test Recall: [illegible]

Late Fusion of Audio & Video Modalities

Video Features

Extracted from facial expression frames using a fine-tuned ResNet50 model.

69%

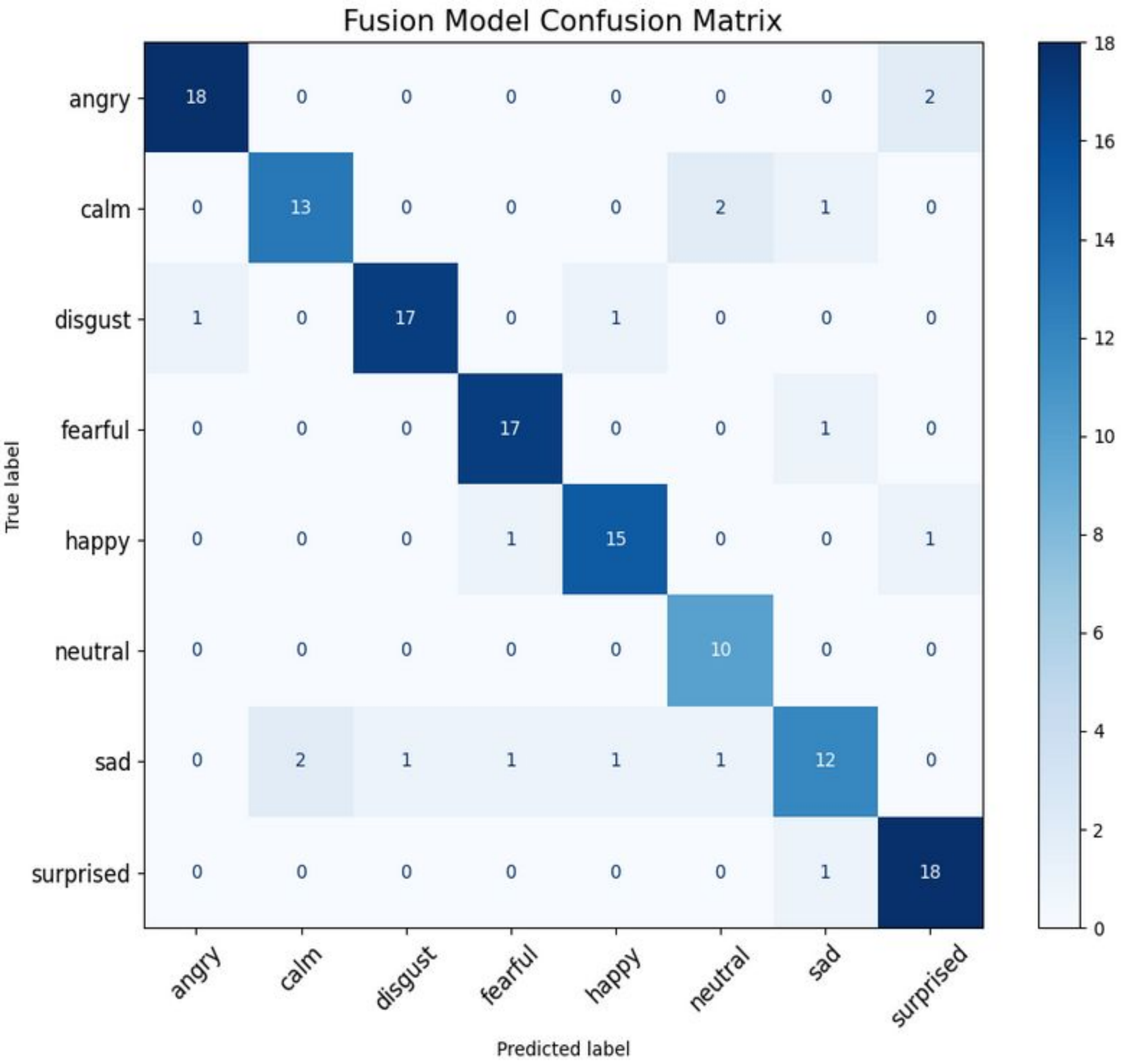
70%
&
76%

Audio Features & spectrogram

Processed MFCCs, waveform images using CNN, BiLSTM and attention layers. and Resnet50 for spectrogram

Fusion Strategy in Multimodal Emotion Recognition

- Three modality-specific architectures applied: **Video, Audio, and Image**
- **Entropy of probabilities** used to assess prediction confidence per modality
- **Weights assigned** using: `weight = trust × confidence`
- **Weighted fusion** computed as: $\sum (\text{weight}_i \times \text{probabilities}_i)$
- **Final emotion selected** using: `emotions[argmax(predictions)]`



Fusion Classification Report:

	precision	recall	f1-score	support
angry	0.95	0.90	0.92	20
calm	0.87	0.81	0.84	16
disgust	0.94	0.89	0.92	19
fearful	0.89	0.94	0.92	18
happy	0.88	0.88	0.88	17
neutral	0.77	1.00	0.87	10
sad	0.80	0.67	0.73	18
surprised	0.86	0.95	0.90	19
accuracy			0.88	137
macro avg	0.87	0.88	0.87	137
weighted avg	0.88	0.88	0.87	137

Novelty: Our Project vs. Research Paper

- **Entropy-weighted dynamic fusion** vs. fixed/basic fusion in the research paper
- Mathematical formula for weighted fusion: `weight = trust × confidence`
- our project uses **Video, Audio, and Image** instead of Text for third modality
- **Adaptive fusion** enables instance-wise reliability assessment

Thank you!