Indian Institite of Information
Tehcnology, Raichur

Dr, DublyaCharya Gyaneswar

VOCALSENSE

iMPLEMENTATION OF RESEARCH PAPER LSTM based Emotion Detection

ADITYA UPENDRA GUPTA
AD24B1003

ANSHIKA AGARWAL
AD24B1007

KARTAVYA GUPTA
AD24B1028

CONTENTS

Introduction

SPEECH EMOTION RECOGNITION (SER)

- AUTOMATICALLY IDENTIFIES SPEAKER'S EMOTIONAL STATE FROM AUDIO
- CRITICAL FOR CUSTOMER ANALYTICS, MENTAL HEALTH, HUMAN-COMPUTER INTERACTION

CHALLENGES:

- SPEAKER VARIABILITY
- LIMITED DATASETS
- TEMPORAL DEPENDENCIES IN SPEECH

OUR SOLUTION:

- DUAL-LAYER LSTM ARCHITECTURE BASED ON YANG ET AL.
- DATASET EXPANSION: REVDAS (1,440) → MIXED DATASETS (11,680 FILES)
- ACHIEVED 83.26% ACCURACY WITH <100MS INFERENCE TIME

Literature Survey - Traditional vs Deep Learning

TRADITIONAL APPROACHES:

- **HAND-CRAFTED FEATURES (MFCCS, PITCH, ENERGY)**
- **CLASSICAL ML: SVMS, GMMS, HMMS**
- **LIMITATIONS: EXTENSIVE FEATURE ENGINEERING, POOR TEMPORAL MODELING**

DEEP LEARNING ERA:

- **CNNS FOR SPATIAL PATTERNS FROM SPECTROGRAMS**
- **RNNS/LSTMS FOR TEMPORAL CONTEXT**
- **HYBRID CNN-LSTM: STATE-OF-THE-ART ON BENCHMARKS**

DUAL-LAYER LSTM (YANG ET AL.):

- **HIERARCHICAL STRUCTURE: LAYER 1 (SHORT-TERM) + LAYER 2 (LONG-TERM PATTERNS)**
- **IMPROVED ACCURACY WHILE MAINTAINING LOW LATENCY (<100 MS)**

Methodology - Preprocessing & Features

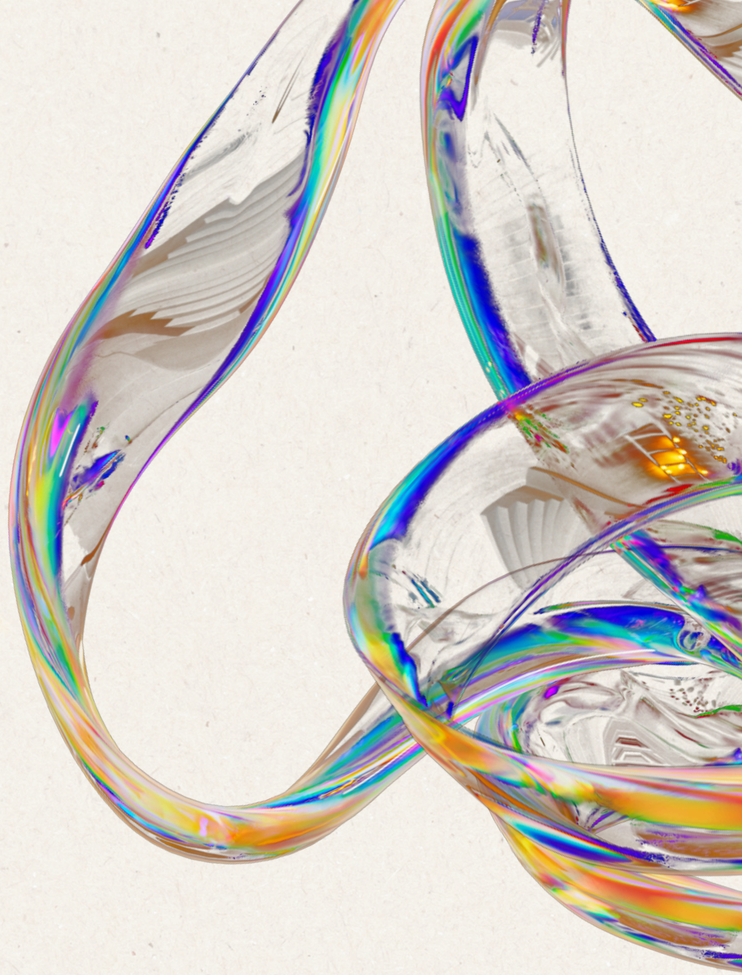
PREPROCESSING:

- **RESAMPLING: 22,050 HZ**
- **MONO CONVERSION**
- **FIXED DURATION: 3.0 SECONDS (TRUNCATION/ZERO-PADDING)**

FEATURE EXTRACTION:

- **MFCCS (40) + DELTA MFCCS → TIMBRE**
- **CHROMA & TONNETZ → HARMONIC CONTENT**
- **MEL-SPECTROGRAMS → FREQUENCY-TIME REPRESENTATION**
- **SPECTRAL CONTRAST → TEXTURE**

NORMALIZATION: STANDARDSCALER FOR STABLE CONVERGENCE



Methodology - Preprocessing & Features

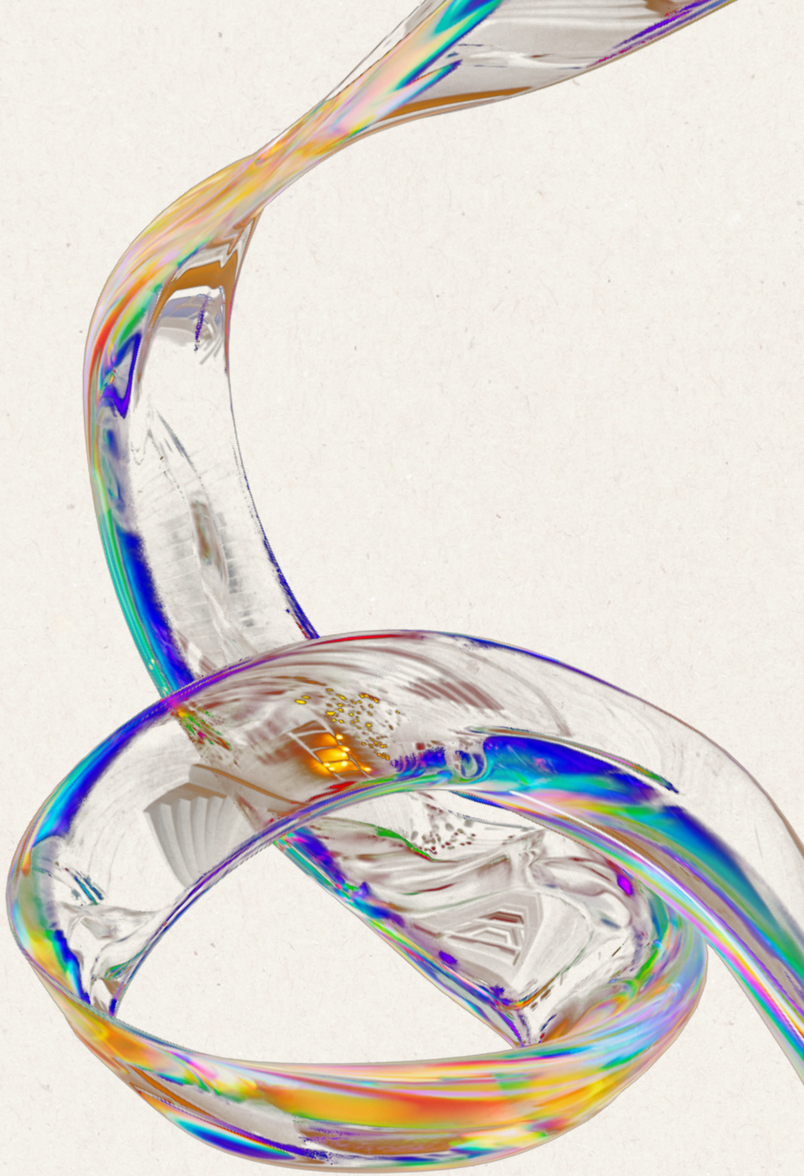
PREPROCESSING:

- **RESAMPLING: 22,050 HZ**
- **MONO CONVERSION**
- **FIXED DURATION: 3.0 SECONDS (TRUNCATION/ZERO-PADDING)**

FEATURE EXTRACTION:

- **MFCCS (40) + DELTA MFCCS → TIMBRE**
- **CHROMA & TONNETZ → HARMONIC CONTENT**
- **MEL-SPECTROGRAMS → FREQUENCY-TIME REPRESENTATION**
- **SPECTRAL CONTRAST → TEXTURE**

NORMALIZATION: STANDARDSCALER FOR STABLE CONVERGENCE



Methodology - Model Architecture

INPUT FEATURES (T, D)

↓

LSTM LAYER 1: 256 UNITS (RETURN_SEQUENCES=TRUE)

↓

BATCH NORMALIZATION + DROPOUT (0.4)

↓

LSTM LAYER 2: 128 UNITS (RETURN_SEQUENCES=FALSE)

↓

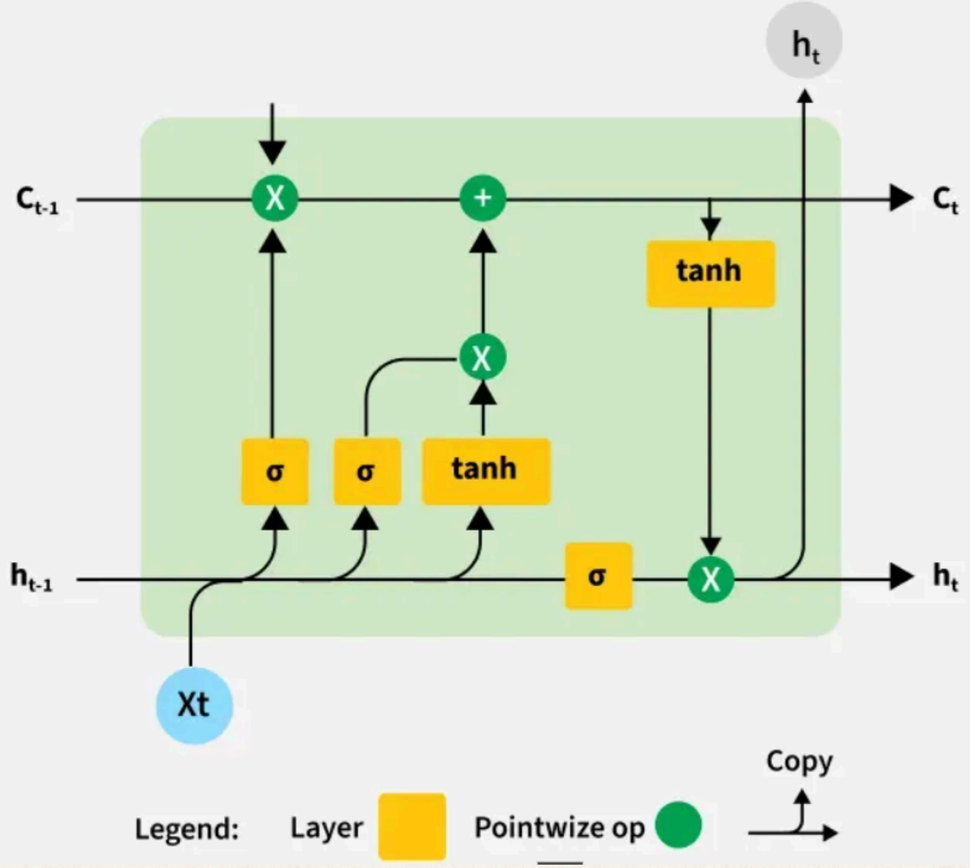
BATCH NORMALIZATION + DROPOUT (0.5)

↓

DENSE LAYER: 256 UNITS (RELU)

↓

OUTPUT: 8 CLASSES (SOFTMAX)



Results - Training Dynamics & Inference

TRAINING CONVERGENCE:

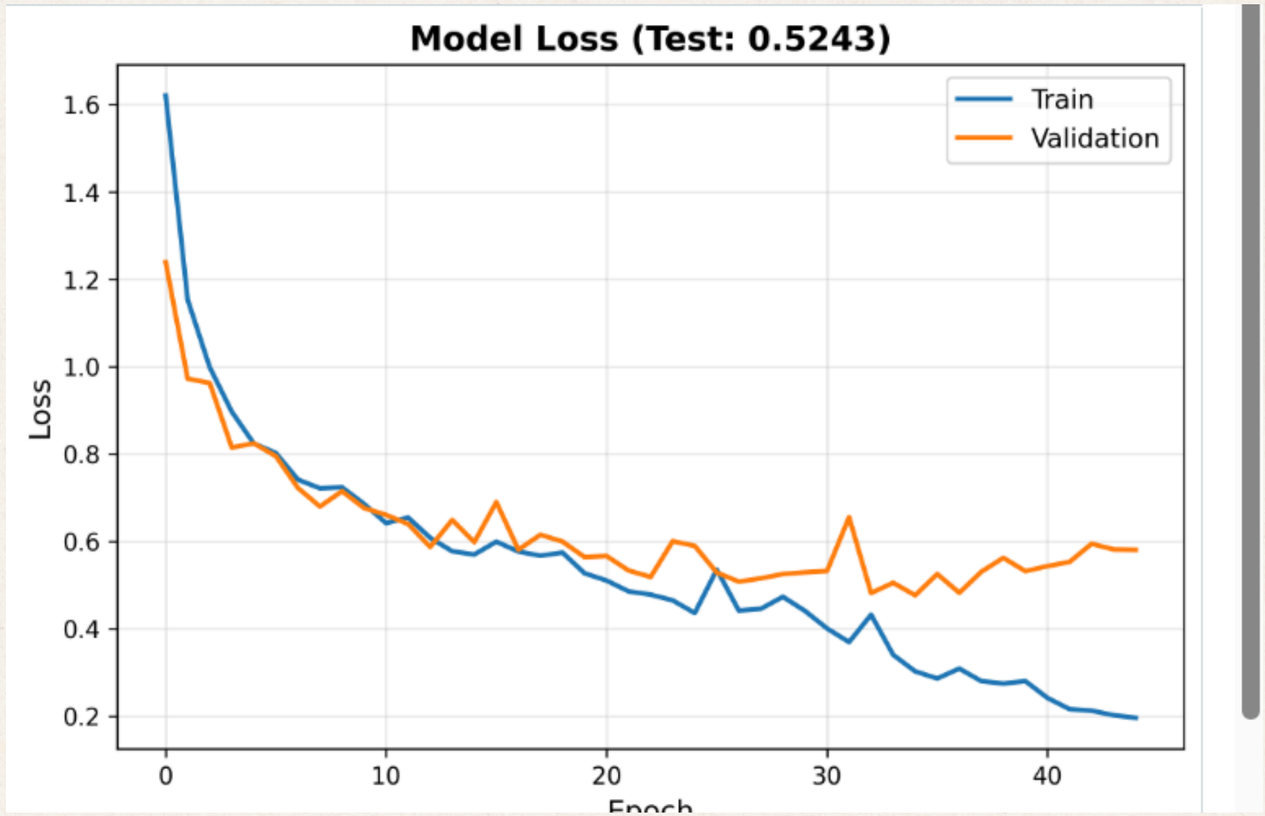
- STABLE TRAINING AND VALIDATION CURVES
- NO SIGNIFICANT OVERFITTING
- REDUCELRONPLATEAU ENABLED FINE-TUNING

INFERENCE PERFORMANCE:

- LATENCY: <100 MS PER 3-SECOND AUDIO CLIP
- HARDWARE: STANDARD CPU (NO GPU REQUIRED)
- SUITABILITY: REAL-TIME APPLICATIONS

DEPLOYMENT:

- STREAMLIT WEB INTERFACE
- FEATURES: AUDIO UPLOAD, PREDICTION, VISUALIZATION (WAVEFORM, MEL-SPECTROGRAM)



Results - Error Analysis

CONFUSION MATRIX INSIGHTS:

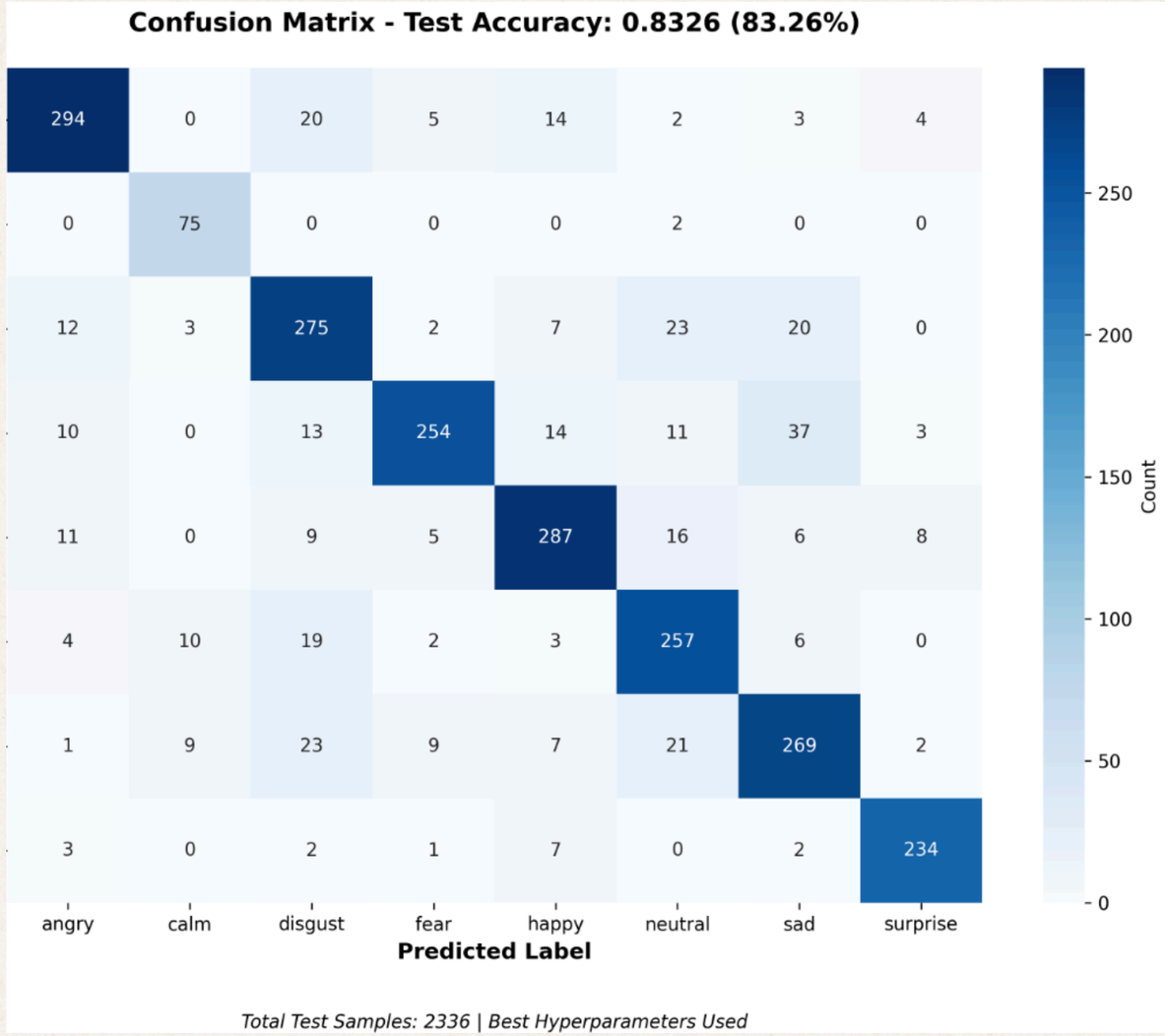
HIGH PERFORMANCE EMOTIONS (>85% ACCURACY):

- ✓ ANGRY, FEARFUL, DISGUST, CALM
- DISTINCTIVE ACOUSTIC CHARACTERISTICS

CHALLENGING EMOTIONS (CONFUSION AREAS):

- × HAPPY ↔ SAD ↔ NEUTRAL
- SIMILAR LOW-AROUSAL ACOUSTIC PROFILES
- OVERLAPPING PITCH AND ENERGY PATTERNS

TAKEAWAY: HIGH-AROUSAL EMOTIONS ARE EASIER TO DISTINGUISH THAN LOW-AROUSAL STATES



Results - Training Dynamics & Inference

TRAINING CONVERGENCE:

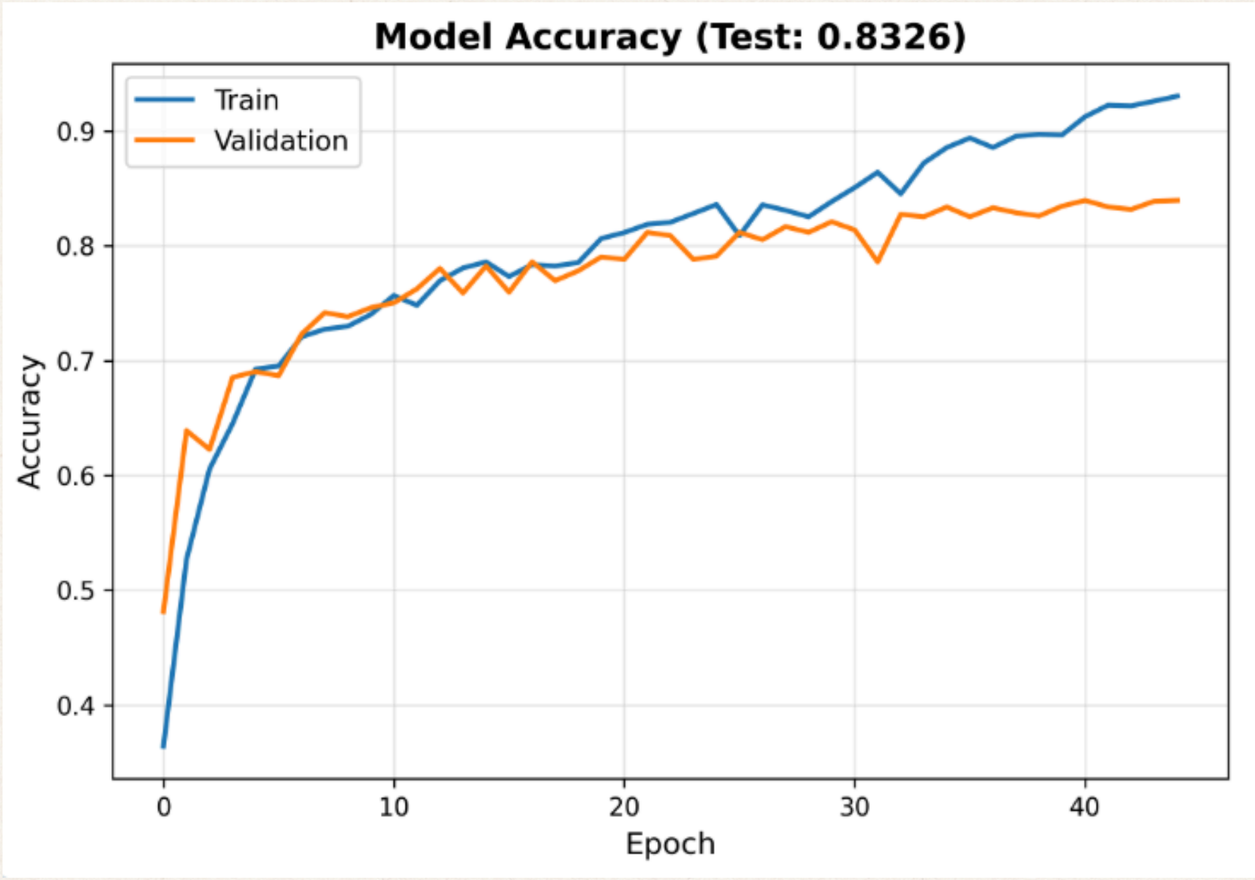
- **STABLE TRAINING AND VALIDATION CURVES**
- **NO SIGNIFICANT OVERFITTING**
- **REDUCELRONPLATEAU ENABLED FINE-TUNING**

INFERENCE PERFORMANCE:

- **LATENCY: <100 MS PER 3-SECOND AUDIO CLIP**
- **HARDWARE: STANDARD CPU (NO GPU REQUIRED)**
- **SUITABILITY: REAL-TIME APPLICATIONS**

DEPLOYMENT:

- **STREAMLIT WEB INTERFACE**
- **FEATURES: AUDIO UPLOAD, PREDICTION, VISUALIZATION (WAVEFORM, MEL-SPECTROGRAM)**



Analysis - Key Insights

1. DATASET IMPACT:
   - EXPANDING FROM 1,440 TO 11,680 FILES = +10% ACCURACY
   - DATA DIVERSITY > MODEL COMPLEXITY
2. ARCHITECTURE EFFECTIVENESS:
   - DUAL-LAYER LSTM BALANCES EFFICIENCY WITH TEMPORAL MODELING
   - HIERARCHICAL LEARNING CAPTURES BOTH LOCAL AND GLOBAL PATTERNS
3. HYPERPARAMETER OPTIMIZATION:
   - DROPOUT (0.4-0.5) CRITICAL FOR PREVENTING OVERFITTING
   - LEARNING RATE 0.001 OPTIMAL FOR CONVERGENCE
   - BATCH SIZE 128 IMPROVED GENERALIZATION
4. FEATURE ENGINEERING:
   - MULTI-FEATURE APPROACH (MFCCS, CHROMA, TONNETZ, SPECTRAL CONTRAST) ENHANCED ROBUSTNESS

Conclusions

PROJECT SUMMARY:

- IMPLEMENTED DUAL-LAYER LSTM-BASED SER SYSTEM
- EXPANDED DATASET FROM 1,440 TO 11,680 SAMPLES ACROSS 8 EMOTIONS
- ACHIEVED 83.26% TEST ACCURACY (10% IMPROVEMENT OVER BASELINE)
- DEPLOYED INTERACTIVE STREAMLIT APPLICATION WITH <100MS INFERENCE

KEY ACHIEVEMENTS: ✓ ROBUST PREPROCESSING AND MULTI-FEATURE EXTRACTION PIPELINE ✓ SYSTEMATIC HYPERPARAMETER OPTIMIZATION (500+ CONFIGURATIONS) ✓ REAL-TIME DEPLOYMENT CAPABILITY ✓ STRONG PERFORMANCE ON HIGH-AROUSAL EMOTIONS

DEMONSTRATION: DUAL-LAYER LSTMS + DIVERSE DATASETS = ROBUST SER PERFORMANCE

References

[1] X. YANG, S. YU, AND W. XU, "IMPROVEMENT AND IMPLEMENTATION OF A SPEECH EMOTION RECOGNITION MODEL BASED ON DUAL-LAYER LSTM"

[2] S. R. LIVINGSTONE AND F. A. RUSSO, "THE RYERSON AUDIO-VISUAL DATABASE OF EMOTIONAL SPEECH AND SONG (RAVDESS)," PLOS ONE, 2018

[3] S. HOCHREITER AND J. SCHMIDHUBER, "LONG SHORT-TERM MEMORY," NEURAL COMPUTATION, VOL. 9, NO. 8, 1997

[4] M. EL AYADI, M. S. KAMEL, AND F. KARRAY, "SURVEY ON SPEECH EMOTION RECOGNITION," PATTERN RECOGNITION, 2011

[5] B. MCFEE ET AL., "LIBROSA: AUDIO AND MUSIC SIGNAL ANALYSIS IN PYTHON," PYTHON IN SCIENCE CONF., 2015

[6] M. ABADI ET AL., "TENSORFLOW: A SYSTEM FOR LARGE-SCALE MACHINE LEARNING," USENIX OSDI, 2016