“Real Time Mental Health Analysis- Well Mind”

Synopsis-II

Submitted in the partial fulfilment of the requirement

for the award of a degree of

**BACHELORS OF SCIENCE**

**IN**

**COMPUTATIONAL STATISTICS & DATA ANALYTICS**

**(2022-2025)**

Submitted to

**DEPARTMENT OF COMPUTATIONAL STATISTICS AND DATA ANALYTICS**

**GURU NANAK DEV UNIVERSITY**

**AMRITSAR**



|  |  |  |
| --- | --- | --- |
| **SUBMITTED TO:** |  | **SUBMITTED BY:** |
| Dr. Sandeep Sharma (HOD) |  | Dipender Singh |
| Dr. Prabhsimran Singh (Coordinator) |  | 17332204549 |

**TABLE OF CONTENTS**

| **S. No** | **Title** | **Page Number** |
| --- | --- | --- |
| 1. | INTRODUCTION | 3-5 |
| 2. | WHY THIS PROJECT? | 6 |
| 3. | OBJECTIVES | 7-8 |
| 4. | FRAMEWORK DEVELOPMENT | 9-12 |
| 5. | PROJECT PROGRESS | 13-15 |
| 6. | EXPLORATORY DATA ANALYSIS | 16-17 |

**INTRODUCTION**

**Overview of the Project**

Mental health is a cornerstone of human well-being, influencing emotional stability, psychological resilience, and social interactions. The global rise in mental health disorders—such as anxiety, depression, and stress—has underscored the need for early detection and intervention to improve quality of life and reduce societal burdens. In India, the prevalence of mental health issues is exacerbated by cultural stigma, limited access to professionals, and reliance on subjective diagnostic methods like clinical assessments and self-reported questionnaires. These traditional approaches are often time-consuming, prone to human bias, and inaccessible, particularly in rural areas where mental health infrastructure is sparse.

The *Real-Time Mental Health Analysis - Well Mind* project addresses these challenges by developing a sophisticated, AI-driven system for real-time emotion detection. The system integrates four data modalities—speech, facial expressions, text, and speech-to-text transcripts—leveraging advanced Machine Learning (ML) and Deep Learning (DL) techniques. Specifically, it employs Convolutional Neural Networks (CNNs) for facial analysis, Long Short-Term Memory (LSTM) networks and Transformer-based models (e.g., Wav2Vec 2.0) for speech processing, Transformer architectures (e.g., RoBERTa) for text and transcript analysis, and Automatic Speech Recognition (ASR) models (e.g., Whisper) for speech-to-text transcription. By fusing these modalities using attention-based and tensor fusion techniques, the system aims to deliver objective, scalable, and accurate mental health assessments, enabling early intervention and personalized recommendations.

The project’s multi-modal approach is designed to capture nuanced emotional cues that single-modality systems often miss. For instance, speech analysis can detect prosodic features like pitch and energy, facial analysis can identify micro-expressions, text analysis can uncover sentiment and context, and speech-to-text transcripts provide additional textual data for deeper emotional insights. This holistic integration enhances diagnostic accuracy and supports real-time applications, such as virtual therapy platforms or self-monitoring tools.

**Motivation**

The motivation for this project stems from the urgent need to address mental health challenges in India, where societal stigma and resource constraints hinder timely diagnosis. According to the World Health Organization (WHO), India faces a significant mental health burden, with over 56 million people affected by depression and 38 million by anxiety disorders as of 2020. Yet, the country has only 0.75 psychiatrists per 100,000 people, compared to the global average of 4.0. Cultural taboos often discourage individuals from seeking help, leading to undiagnosed conditions that escalate into severe mental health crises.

Traditional diagnostic methods rely heavily on subjective inputs, such as patient self-reports or clinician observations, which can be inconsistent and biased. Moreover, these methods are not scalable to meet the needs of India’s 1.4 billion population, particularly in rural areas with limited healthcare access. Technology-driven solutions, particularly those leveraging AI and Data Science, offer a promising avenue to bridge this gap. By automating emotion detection through multi-modal analysis, the *Well Mind* system aims to:

* **Enable Early Detection**: Identify emotional distress in real-time, facilitating timely interventions to prevent escalation.
* **Reduce Bias**: Use objective, AI-driven analysis to minimize human subjectivity in mental health assessments.
* **Enhance Accessibility**: Deliver a scalable solution accessible via web and mobile platforms, reaching underserved populations.
* **Provide Holistic Insights**: Integrate speech, facial, text, and speech-to-text data to capture comprehensive emotional states, improving diagnostic robustness.

This project aligns with the growing field of affective computing, which combines Data Science, psychology, and AI to understand human emotions. By addressing cultural and infrastructural barriers, it seeks to foster a stigma-free environment for mental health assessment in India.

**Project Objectives**

The primary goal is to develop a real-time, AI-powered mental health detection system that integrates multi-modal data for accurate emotion classification. The system is designed to operate with low latency, ensuring applicability in live scenarios like virtual therapy or personal monitoring. The specific objectives are:

* **Analyze Acoustic Features**: Extract prosodic features (e.g., Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, zero-crossing rate) from speech signals to classify emotional states using advanced models like CNN-LSTM or Wav2Vec 2.0.
* **Detect Facial Expressions**: Apply CNN-based architectures (e.g., VGG-16, ResNet-50) to identify emotions from facial landmarks and micro-expressions in real-time video feeds, achieving high accuracy (>85% F1-score).
* **Perform Sentiment Analysis**: Process text inputs and speech-to-text transcripts using Transformer-based models (e.g., RoBERTa, DistilBERT) to extract emotional context and sentiment polarity.
* **Implement Automatic Speech Recognition (ASR)**: Transcribe real-time speech into text using state-of-the-art ASR models (e.g., DeepSpeech, Whisper), enabling sentiment analysis on verbal inputs.
* **Fuse Multi-Modal Data**: Employ advanced fusion techniques, such as multi-head attention or tensor fusion, to integrate predictions across modalities, enhancing classification robustness and accuracy.
* **Optimize for Real-Time Inference**: Develop low-latency inference pipelines using model optimization techniques (e.g., ONNX, TensorRT) for deployment in real-world applications.
* **Deliver Actionable Insights**: Generate comprehensive mental health status reports with visualizations (e.g., emotional trend graphs) and recommend interventions based on detected emotional states.

**Expected Outcome**

The project aims to deliver:

* A robust multi-modal AI model achieving >85% F1-score and AUC-ROC across speech, facial, text, and speech-to-text modalities, validated through rigorous evaluation metrics.
* A scalable, user-friendly application deployed via Flask or FastAPI, featuring real-time interaction and visualization dashboards for emotional trends.
* A system that supports healthcare professionals and individuals with actionable insights, reducing diagnostic delays and promoting proactive mental health management.
* A contribution to affective computing and mental health diagnostics, addressing cultural barriers and improving accessibility in India.

This project represents a significant advancement in AI-driven mental health solutions, leveraging Data Science to enable early intervention and personalized care.

**WHY THIS PROJECT?**

Mental health remains a critical yet under-addressed issue in India, where cultural norms and societal stigma discourage open discussions about emotional well-being. According to the National Mental Health Survey (2015-16), nearly 15% of Indian adults require active mental health interventions, yet stigma leads to fewer than 30% seeking professional help. The shortage of mental health professionals—approximately 9,000 psychiatrists for a population of 1.4 billion—combined with high therapy costs and limited rural infrastructure, exacerbates the problem. Mental health issues are often dismissed as transient or a sign of weakness, delaying diagnosis and treatment.

Traditional diagnostic methods, such as clinical interviews and questionnaires (e.g., PHQ-9, GAD-7), rely on subjective inputs, which are prone to biases and inconsistencies. These methods are neither scalable nor accessible to India’s diverse population, particularly in remote areas with limited healthcare facilities. The advent of AI and Data Science offers a transformative opportunity to address these challenges through automated, objective, and scalable solutions.

The *Well Mind* project leverages advanced ML and DL techniques to develop a real-time mental health detection system that analyzes emotions through four modalities: speech, facial expressions, text, and speech-to-text transcripts. By integrating these data streams, the system captures a comprehensive view of emotional states, overcoming the limitations of single-modality approaches. For example:

* **Speech Analysis**: Detects prosodic features like pitch, intonation, and speech rate, which are correlated with emotional states like anxiety or sadness.
* **Facial Analysis**: Identifies micro-expressions and facial muscle movements (e.g., AU12 for smiling, AU4 for frowning) using CNNs, aligned with Ekman’s Facial Action Coding System (FACS).
* **Text Analysis**: Extracts sentiment and context from user inputs and ASR transcripts, capturing linguistic cues of emotional distress.
* **Speech-to-Text**: Converts spoken inputs into text for NLP-based sentiment analysis, enhancing the system’s ability to process verbal emotional cues.

This multi-modal approach ensures robustness and accuracy, making the system a viable first point of contact for individuals hesitant to seek traditional help. By deploying the system on accessible platforms (web, mobile), it addresses infrastructural barriers, particularly in rural India. Beyond technological innovation, the project aims to foster a cultural shift toward open mental health discussions, reducing stigma and promoting psychological well-being.

**OBJECTIVES**

**Primary Objectives**

1. **Develop a Multi-Modal Emotion Recognition System**
   * Design an AI system to process real-time speech, facial, text, and speech-to-text data using advanced architectures (e.g., CNNs, LSTMs, Transformers).
   * Achieve high classification performance (>85% F1-score, >0.9 AUC-ROC) through end-to-end training, fine-tuning, and cross-modal integration.
   * Implement robust preprocessing pipelines for feature extraction, including MFCCs for speech, facial landmarks for images, and word embeddings for text/transcripts.
2. **Enable Real-Time Mental Health Monitoring**
   * Develop low-latency inference pipelines using optimized models (e.g., quantized CNNs, ONNX-converted Transformers) for real-time emotion detection.
   * Support applications like virtual therapy or self-monitoring apps with seamless multi-modal processing and minimal latency (<500ms per inference).
   * Ensure robustness to noisy inputs (e.g., low-light images, background noise in audio) through advanced preprocessing and denoising techniques.
3. **Enhance Accuracy via Multi-Modal Fusion**
   * Implement advanced fusion techniques, such as multi-head attention, tensor fusion, or hybrid fusion, to integrate predictions from speech, facial, text, and speech-to-text modalities.
   * Optimize fusion weights using gradient-based methods or Bayesian optimization to maximize cross-modal synergy.
   * Mitigate modality-specific limitations (e.g., speech ambiguity, facial occlusion) to achieve robust emotional classification.

**Secondary Objectives**

1. **Develop an Interactive Interface**
   * Create a web-based interface using Flask or FastAPI, enabling real-time input processing (speech, video, text) and visualization of emotional trends (e.g., time-series plots, heatmaps).
   * Ensure cross-platform compatibility (web, iOS, Android) with responsive design and low computational overhead.
   * Incorporate user feedback loops for iterative model improvement and personalized recommendations.
2. **Leverage State-of-the-Art Models**
   * Utilize pre-trained models (e.g., VGG-16 for facial analysis, Wav2Vec 2.0 for speech, RoBERTa for text/transcripts) with transfer learning to reduce training time.
   * Fine-tune models using grid search or Bayesian optimization for hyperparameters (e.g., learning rate, batch size, dropout rate).
   * Implement regularization techniques (e.g., L2 regularization, dropout) to prevent overfitting on small datasets.
3. **Ensure Scalability and Robustness**
   * Deploy models on cloud platforms (e.g., AWS EC2, GCP Cloud Run) with load balancing for large-scale usage.
   * Optimize inference with ONNX or TensorRT to achieve real-time performance on resource-constrained devices.
   * Develop robust preprocessing pipelines to handle real-world data variability (e.g., accents in speech, diverse lighting in images).

These objectives aim to deliver a technically advanced, scalable system for real-time mental health monitoring, suitable for academic research and practical deployment.

**FRAMEWORK DEVELOPMENT**

The *Well Mind* system is a sophisticated, multi-modal framework designed for real-time emotion detection and mental health analysis. It integrates four data streams—speech, facial expressions, text, and speech-to-text transcripts—processed through a pipeline of feature extraction, model training, fusion, and deployment. Below is a detailed breakdown of the framework’s components.

**Data Collection and Preprocessing**

**Facial Recognition**

* **Dataset**: A custom dataset of 3,000+ grayscale images (48x48 pixels) manually labeled for eight emotions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, Surprise). Images are captured in controlled settings to ensure high quality and accurate labeling.
* **Preprocessing**:
  + **Normalization**: Scale pixel intensities to [0, 1] to standardize input for CNNs.
  + **Data Augmentation**: Apply transformations (rotation ±15°, flipping, brightness adjustment ±20%, shear ±10°) to enhance model robustness to real-world variations (e.g., lighting, pose).
  + **Facial Landmarks**: Extract 68 facial landmarks using DLib or MTCNN to capture muscle movements (e.g., Action Units per FACS). Compute landmark distances and angles as additional features.
  + **Dimensionality Reduction**: Apply Principal Component Analysis (PCA) to reduce feature dimensionality while retaining 95% variance.
* **Feature Extraction**: Use pre-trained CNNs (e.g., VGG-16, ResNet-50) to extract high-level features from convolutional layers, supplemented by landmark-based features for micro-expression analysis.

**Speech Emotion Recognition**

* **Datasets**: Curated from RAVDESS (1,440 samples), CREMA-D (7,442 samples), IEMOCAP (5,531 samples), and TESS (2,800 samples), totaling ~17,000 labeled audio samples across eight emotions.
* **Preprocessing**:
  + **Feature Extraction**: Extract 40 MFCCs, 12 chroma features, spectral contrast, pitch, energy, and zero-crossing rate using Librosa. Generate log-mel spectrograms for time-frequency analysis.
  + **Noise Reduction**: Apply spectral gating and Wiener filtering to mitigate background noise, ensuring robust feature extraction in real-world environments.
  + **Normalization**: Standardize features to zero mean and unit variance to stabilize model training.
* **Feature Engineering**: Compute temporal statistics (mean, variance, skewness) of acoustic features to capture prosodic patterns. Use delta and delta-delta coefficients to model temporal dynamics.

**Speech-to-Text Transcription**

* **Pipeline**: Implement ASR using OpenAI Whisper or Mozilla DeepSpeech for real-time transcription of audio inputs. Whisper, a Transformer-based model, is preferred for its robustness to diverse accents and noisy environments.
* **Preprocessing**:
  + **Audio Preprocessing**: Resample audio to 16 kHz, apply noise reduction, and normalize amplitude.
  + **Text Preprocessing**: Tokenize transcripts using spaCy, remove stop-words, and apply lemmatization to standardize text. Perform Named Entity Recognition (NER) to identify contextual entities (e.g., names, locations).
  + **Feature Extraction**: Generate word embeddings using RoBERTa or DistilBERT, capturing semantic and emotional context. Compute sentiment scores using NLTK’s VADER and polarity/subjectivity with TextBlob.
* **Performance Metrics**: Evaluate ASR using Word Error Rate (WER), targeting <15% on clean audio. Assess sentiment analysis accuracy (>80% F1-score) on transcribed text.

**Text Sentiment Analysis**

* **Dataset**: Real-time user inputs and ASR transcripts (~2,000 samples), supplemented with public sentiment datasets (e.g., SST-2, IMDb).
* **Preprocessing**:
  + **Tokenization**: Use spaCy for word and sentence tokenization, preserving contextual structure.
  + **Feature Extraction**: Generate contextual embeddings with RoBERTa, capturing long-range dependencies. Compute readability scores (Flesch-Kincaid) to assess text complexity.
  + **NER and POS Tagging**: Identify entities and parts of speech to enhance emotional context understanding.
* **Feature Engineering**: Extract n-grams (unigrams, bigrams) and sentiment-specific features (e.g., polarity, subjectivity) to complement embeddings.

**Model Development**

**Human Detection**

* **Model**: Use YOLOv5 or MTCNN for real-time human detection in video feeds. YOLOv5 is preferred for its speed (30 FPS on GPU) and accuracy in detecting faces under occlusion.
* **Training**: Fine-tune on COCO dataset with additional face-specific annotations to improve detection in diverse settings.
* **Output**: Bounding boxes and confidence scores to ensure only relevant subjects are analyzed.

**Emotion Recognition**

* **Facial Recognition**:
  + **Model**: Fine-tune VGG-16 or ResNet-50 with ImageNet weights, adding dense layers for emotion classification (8 classes). Use softmax activation for multi-class output.
  + **Training**: Optimize with AdamW (learning rate 1e-4, weight decay 1e-2) and categorical cross-entropy loss. Apply dropout (0.5) and batch normalization to prevent overfitting.
  + **Augmentation**: Use real-time augmentation with Albumentations to simulate diverse lighting and pose conditions.
* **Speech Emotion Recognition**:
  + **Model**: Implement a hybrid CNN-LSTM architecture to capture spatial and temporal features. Alternatively, use Wav2Vec 2.0, a Transformer-based model pre-trained on large-scale audio data.
  + **Training**: Optimize with Adam (learning rate 1e-5) and focal loss to address class imbalance. Use early stopping based on validation loss.
  + **Feature Input**: Feed log-mel spectrograms and MFCCs, with sequence length padded to 3 seconds for consistency.
* **Speech-to-Text Sentiment Analysis**:
  + **Model**: Fine-tune RoBERTa or DistilBERT for sentiment classification on transcripts. Use multi-task learning to predict both sentiment (positive, negative, neutral) and emotion (aligned with speech/facial classes).
  + **Training**: Optimize with AdamW (learning rate 2e-5) and cross-entropy loss. Apply gradient clipping to stabilize training.
  + **Input**: Tokenized transcripts with maximum sequence length of 512 tokens, padded or truncated as needed.
* **Text Sentiment Analysis**:
  + **Model**: Use RoBERTa with attention layers for context-aware sentiment classification. Supplement with VADER for fine-grained sentiment intensity.
  + **Training**: Fine-tune on mixed datasets (SST-2, custom inputs), using mixed-precision training for efficiency.
  + **Input**: Contextual embeddings from RoBERTa, augmented with n-gram features and readability scores.

**Multi-Modal Fusion**

* **Techniques**:
  + **Early Fusion**: Concatenate feature vectors from all modalities (e.g., CNN features, MFCCs, RoBERTa embeddings) and feed into a dense layer.
  + **Decision-Level Fusion**: Combine modality-specific predictions using weighted averaging, with weights optimized via grid search.
  + **Attention-Based Fusion**: Implement multi-head attention to dynamically weight modality contributions based on input context.
  + **Tensor Fusion**: Use tensor outer products to capture pairwise interactions between modalities, increasing model expressiveness.
* **Training**: Optimize fusion layer with Adam (learning rate 1e-4) and cross-entropy loss. Use L1 regularization to prevent overfitting.
* **Evaluation**: Assess fusion performance with F1-score, AUC-ROC, and confusion matrices, targeting a 5-10% improvement over single-modality models.

**Evaluation Metrics**

* **Classification Metrics**: Accuracy, precision, recall, F1-score, AUC-ROC, Cohen’s Kappa for multi-class performance.
* **Regression Metrics**: Mean Absolute Error (MAE) for continuous sentiment scores (e.g., VADER polarity).
* **Cross-Validation**: Use 5-fold cross-validation to ensure generalizability. Compute macro- and micro-averaged metrics to handle class imbalance.
* **Error Analysis**: Generate confusion matrices and per-class error rates to identify misclassification patterns.

**Deployment**

* **Platform**: Deploy models using Flask or FastAPI on cloud infrastructure (AWS EC2, GCP Cloud Run) with auto-scaling for high throughput.
* **Optimization**: Convert models to ONNX or TensorRT formats for low-latency inference (<200ms per sample on GPU). Use mixed-precision inference to reduce memory footprint.
* **Interface**: Develop a web-based interface with real-time visualizations (e.g., Plotly for time-series emotional trends, heatmaps for feature importance).
* **Scalability**: Implement load balancing and caching to handle concurrent users. Use Docker containers for consistent deployment across environments.

This framework delivers a technically sophisticated, scalable solution for real-time mental health analysis, integrating advanced Data Science methodologies and robust preprocessing pipelines.

**PROJECT PROGRESS**

Over the past five months, substantial progress has been made in developing the *Well Mind* system, with a focus on technical implementation, experimentation, and dataset curation. Below is a detailed overview of achievements and challenges.

**Technical Skill Development**

* **Frameworks**: Mastered TensorFlow 2.x and PyTorch 1.x for DL model development, focusing on CNNs, LSTMs, and Transformers. Implemented custom layers for attention-based fusion and feature extraction.
* **ASR Libraries**: Explored SpeechRecognition, DeepSpeech, and OpenAI Whisper for speech-to-text transcription. Achieved preliminary WER of 12-15% on clean audio using Whisper.
* **NLP Tools**: Proficient in spaCy, NLTK, TextBlob, and transformers (Hugging Face) for text and transcript processing.
* **Visualization**: Developed expertise in Plotly, Seaborn, and Matplotlib for EDA and real-time dashboards.

**Facial Emotion Recognition**

* **Dataset**: Transitioned from FER-2013 (Kaggle) to a custom dataset due to noisy labels and low resolution (28x28 pixels). Collected 3,000+ grayscale images (48x48 pixels) across eight emotions, manually labeled with inter-rater agreement (Cohen’s Kappa >0.8).
* **Preprocessing**: Applied normalization, augmentation (rotation, shear, zoom), and landmark extraction using DLib. Reduced feature dimensionality with PCA, retaining 95% variance.
* **Model**: Fine-tuned VGG-16 with ImageNet weights, achieving 78% validation accuracy and 0.85 AUC-ROC after 50 epochs. Experimented with ResNet-50, yielding marginal improvements (80% accuracy).
* **Challenges**: Addressed class imbalance (e.g., Contempt: 250 images, Happiness: 800 images) using SMOTE and weighted loss functions. Mitigated overfitting with dropout (0.5) and L2 regularization (1e-2).

**Speech Emotion Recognition**

* **Dataset**: Curated 17,000+ audio samples from RAVDESS, CREMA-D, IEMOCAP, and TESS. Ensured balanced representation across emotions.
* **Preprocessing**: Extracted 40 MFCCs, 12 chroma features, pitch, and energy using Librosa. Generated log-mel spectrograms with 128 mel bins. Applied spectral gating to reduce noise.
* **Model**: Trained a CNN-LSTM model (2 CNN layers, 1 LSTM layer) with 73% F1-score on RAVDESS test set. Experimented with Wav2Vec 2.0, achieving 76% F1-score after fine-tuning on 10,000 samples.
* **Challenges**: Handled variable audio lengths by padding to 3 seconds. Addressed noisy inputs with Wiener filtering, improving model robustness by 5%.

**Speech-to-Text Transcription**

* **Pipeline**: Implemented OpenAI Whisper for ASR, achieving WER of 12% on clean audio and 20% on noisy samples. Integrated transcripts with RoBERTa for sentiment analysis, yielding 82% F1-score on binary sentiment tasks.
* **Preprocessing**: Resampled audio to 16 kHz, applied noise reduction, and normalized amplitude. Tokenized transcripts with spaCy, achieving 95% tokenization accuracy.
* **Challenges**: Improved WER on accented speech by fine-tuning Whisper on a subset of IEMOCAP. Addressed transcription errors with post-processing rules (e.g., spell-checking).

**Text Sentiment Analysis**

* **Dataset**: Combined 2,000+ real-time inputs and ASR transcripts with public datasets (SST-2, IMDb). Ensured diversity in text length and sentiment.
* **Preprocessing**: Applied tokenization, lemmatization, and NER with spaCy. Generated RoBERTa embeddings with sequence length of 512 tokens.
* **Model**: Fine-tuned RoBERTa for sentiment classification, achieving 87% F1-score and 0.90 AUC-ROC. Integrated VADER for sentiment intensity, with correlation coefficient of 0.85 with ground truth.
* **Challenges**: Handled ambiguous text (e.g., sarcasm) by incorporating context-aware embeddings and multi-task learning.

**Multi-Modal Fusion**

* **Experiments**: Tested early fusion (feature concatenation) and decision-level fusion (weighted averaging), achieving 80% and 82% accuracy, respectively. Implemented multi-head attention, improving accuracy to 85%.
* **Optimization**: Tuned fusion weights using Bayesian optimization, reducing validation loss by 10%. Explored tensor fusion for pairwise modality interactions, pending further evaluation.
* **Challenges**: Addressed modality misalignment (e.g., speech-text latency) by synchronizing inputs using timestamp-based buffering.

**Research and Domain Knowledge**

* **Literature Review**: Analyzed 20+ papers on affective computing, focusing on multi-modal fusion, ASR integration, and emotion recognition. Key references include Ekman’s emotion model and recent advances in Transformer architectures.
* **Psychological Insights**: Studied FACS for facial analysis, prosodic cues for speech, and linguistic markers for text to align AI predictions with clinical standards.
* **Challenges**: Balanced technical complexity with psychological validity, ensuring model outputs are interpretable by healthcare professionals.

**Future Steps**

* Refine fusion techniques with hybrid attention-tensor models.
* Expand custom dataset to 5,000+ images and 20,000+ audio samples.
* Optimize inference pipelines for edge devices (e.g., mobile phones) using TensorRT.
* Conduct user testing to validate interface usability and recommendation accuracy.

This progress reflects a strong foundation for a robust, multi-modal mental health detection system, with ongoing efforts to enhance performance and scalability.

**EXPLORATORY DATA ANALYSIS**

EDA is critical for understanding data characteristics, identifying biases, and informing model design. The *Well Mind* system’s multi-modal dataset is analyzed across four modalities: facial images, speech, speech-to-text transcripts, and text.

**Facial Data**

* **Dataset**: 3,000+ grayscale images (48x48 pixels) manually labeled for eight emotions. Inter-rater agreement (Cohen’s Kappa = 0.82) ensures label reliability.
* **Visualizations**:
  + **t-SNE Embeddings**: Visualize high-dimensional CNN features, revealing clear clusters for Happiness and Sadness but overlap for Contempt and Neutral.
  + **Class Distribution**: Bar chart shows imbalance (e.g., Happiness: 800 images, Contempt: 250 images), necessitating SMOTE or augmentation.
  + **Sample Images**: Grid of representative faces confirms expression clarity and lighting consistency.
  + **Feature Correlation**: Heatmap of landmark distances (e.g., mouth width, eye aperture) shows high correlation (>0.7) for Happiness and Sadness features.
* **Statistical Analysis**: Compute class-wise mean and variance of pixel intensities. Apply Shapiro-Wilk test to confirm normality (p > 0.05).
* **Preprocessing**: Normalize images, apply augmentation, and reduce dimensionality with PCA (95% variance retained).

**Speech Data**

* **Dataset**: 17,000+ audio samples from RAVDESS, CREMA-D, IEMOCAP, and TESS, covering eight emotions.
* **Visualizations**:
  + **Spectrograms**: Log-mel spectrograms (128 mel bins) highlight frequency patterns for emotions (e.g., high energy for Anger).
  + **Feature Correlation Matrix**: Heatmap shows MFCCs and spectral contrast correlation (>0.65), suggesting feature selection.
  + **Waveform Analysis**: Time-domain plots reveal amplitude variations for emotional intensity.
* **Statistical Analysis**: Compute skewness and kurtosis of MFCCs to capture prosodic variations. Apply ANOVA to confirm feature significance (p < 0.01).
* **Preprocessing**: Extract 40 MFCCs, 12 chroma features, and pitch. Apply noise reduction and normalization.

**Speech-to-Text Data**

* **Dataset**: 2,000+ transcripts from Whisper ASR on sample audio inputs.
* **Visualizations**:
  + **Word Clouds**: Highlight frequent emotional keywords (e.g., “stress,” “happy”) across transcripts.
  + **Sentiment Score Distribution**: Histograms of VADER scores show balanced positive/negative sentiment (mean polarity = 0.1).
  + **Token Length Analysis**: Plot distribution of transcript lengths (mean = 20 tokens).
* **Statistical Analysis**: Compute word frequency and TF-IDF scores to identify significant terms. Apply Mann-Whitney U test to compare sentiment distributions (p < 0.05).
* **Preprocessing**: Tokenize with spaCy, remove stop-words, and lemmatize. Generate RoBERTa embeddings for downstream analysis.

**Text Data**

* **Dataset**: 2,000+ real-time inputs and ASR transcripts, supplemented with SST-2 and IMDb datasets.
* **Visualizations**:
  + **Word Embeddings**: t-SNE visualization of RoBERTa embeddings shows semantic clustering by sentiment.
  + **Sentiment Histograms**: Polarity and subjectivity distributions from TextBlob (mean polarity = 0.15, subjectivity = 0.5).
  + **N-Gram Analysis**: Bar chart of top bigrams (e.g., “feel sad,” “very happy”) highlights emotional phrases.
* **Statistical Analysis**: Compute cosine similarity between embeddings to assess semantic similarity. Apply chi-square test to validate sentiment distribution (p < 0.01).
* **Preprocessing**: Tokenize, lemmatize, and apply NER with spaCy. Normalize text for model compatibility.

**Observations**

* **Class Imbalance**: Facial dataset under-represents Contempt and Surprise, requiring oversampling or augmentation.
* **Feature Redundancy**: High correlation in speech features (MFCCs, spectral contrast) suggests PCA or feature selection.
* **ASR Performance**: Whisper achieves low WER (12%) on clean audio but struggles with noise (20% WER), necessitating advanced denoising.
* **Text Ambiguity**: Sarcasm and context-dependent phrases challenge sentiment analysis, requiring multi-task learning.
* **Model Readiness**: Balanced datasets, robust preprocessing, and statistical validation ensure data suitability for training.

**Detailed code, visualizations, and statistical analyses are available in the EDA.ipynb file on the GitHub repository.**