Proposal Ethical AI for Emotional Health: Transparent Depression Prediction

Mentor: Dr Nurfazrina Mohd Zamry

Introduction:

Depression affects 280 million people globally, yet half of cases go undetected due to subjective diagnostic tools and stigma. AI offers a solution but faces ethical challenges like opacity ("black-box" models), bias, and privacy concerns. This project proposes an interpretable, multimodal AI framework to predict depression transparently and equitably.

Proposed Solution:

- Multimodal Analysis: Combines speech (vocal biomarkers), text (linguistic cues), and behavioral data (sleep/smartphone patterns) for comprehensive assessment.
- Explainability: Uses SHAP/LIME to provide clinician-friendly explanations (e.g., "Vocal monotony +22% risk").
- Ethical Safeguards: Implements adversarial debiasing to reduce bias and federated learning with differential privacy (ε=0.5) for data security.
- Clinical Integration: Delivers outputs via an interactive dashboard aligned with DSM-5 criteria.

Methodology:

1. Data: DAIC-WOZ (speech), MODMA (multimodal), and DepSign (behavioral) datasets.
2. Models:
   - Speech: CNN-BiLSTM with layer-wise relevance propagation.
   - Text: MentalBERT (fine-tuned BERT) with SHAP analysis.
   - Behavior: Time-series transformers for anomaly detection.
3. Evaluation: AUC-ROC, fairness metrics (Disparate Impact Ratio), and clinician intuitiveness scores.

Impact:

- Targets 85–92% accuracy, surpassing unimodal approaches.
- Aims to reduce diagnostic delays (currently 9.2 years) and prevent 23% of severe cases through early detection.
- Guides ethical AI deployment in healthcare policy.

Feasibility:

Uses Python/PyTorch, federated learning (PySyft), and cloud GPUs for scalable development.

Conclusion:

This framework bridges AI innovation with ethical mental healthcare, prioritizing transparency, fairness, and actionable clinical insights.

References (Key):

1. WHO (2021). Depression fact sheet.
2. Lundberg & Lee (2017). SHAP for model interpretability. NeurIPS.
3. Obermeyer et al. (2019). Bias in healthcare algorithms. Science.