~ Group BlaBlaBla ~

Ethical AI for Emotional Health: Transparent Depression Prediction

Introduction

Depression is a critical **mental health challenge**, yet underdiagnosed due to stigma, subjective methods and healthcare inequality. Our research explores the design of an **ethical**, **interpretable** and **multimodal AI framework** to support early depression prediction. This framework integrates data from speech, text and behavior while embedding **fairness**, **transparency** and **privacy** as core principles to address critical limitations in current AI approaches.

Problem Statement

Current diagnosis methods like **self-report questionnaires** are vulnerable to **stigma**, **cultural bias** and **underreporting**. While AI has shown promise in predicting depression, **real-world adoption is limited** due to:

- **Lack of interpretability** ("black-box" models)
- **Demographic biases** in predictions
- **Privacy** and **ethical concerns** in sensitive health data use

Proposed Solution

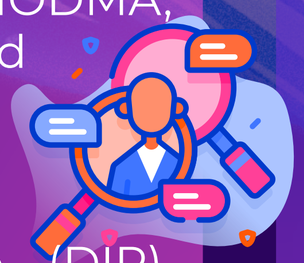
We propose a conceptual AI framework that combines:

- **Speech:** Prosodic features via CNN-BiLSTM (e.g., pitch variability)
- **Text:** Linguistic markers via MentalBERT (e.g., self-focus)
- **Behavior:** Anomalies in phone usage/sleep using time-series transformers

All predictions are explained using **SHAP** and **LIME**. Fairness is addressed via **adversarial debiasing**. Privacy is preserved with **federated learning** and **differential privacy**.

AI Methodology

- **Multimodal Fusion**: Early fusion of features
- **Ethical AI:** Bias auditing, privacy-preserving federated training
- **Tools (Research Basis):** Python, PyTorch, MentalBERT, SHAP, Captum, PySyft
- **Datasets Reviewed:** DAIC-WOZ, MODMA, DepSign, Depression Deconstructed



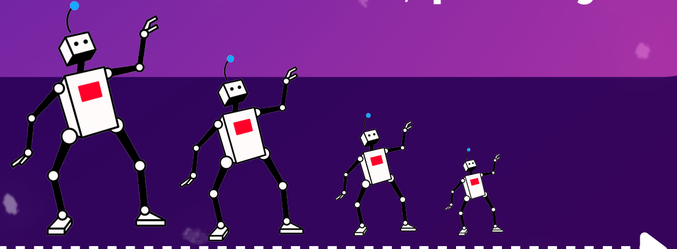
Evaluation metrics proposed:

- **Technical:** AUC-ROC, F1-score
- **Ethical:** Disparate Impact Ratio (DIR), Explanation Fidelity
- **Clinical:** Depression Severity Correlation (DSC)

ESG Impact (AI for Humanity)

- **Environmental:** Reduced compute via federated training
- **Social:** Scalable framework designed for underserved populations
- **Governance:** Aligned with GDPR, HIPAA, and WHO mental health priorities

We aim to ensure AI serves humanity. Not just with accuracy but with **fairness**, **privacy** and **transparency**.



Expected Otcomes

- A conceptual AI model achieving **85–92% accuracy** (based on literature)
- Actionable **clinical insights** (e.g., “vocal monotony increases risk by +22%”)
- **Ethical deployment** guidelines for mental health AI
- **Policy recommendations** for transparent, fair, and privacy-respecting AI in healthcare

Although no working system exists, our research lays the foundation for **future AI tools** that clinicians can trust and patients can benefit from.

Mentor: Dr. Nurfazrina Mohd Zamry

Members: Cornelia Lim Zhi Xuan / Lim En Dhong / Ng Jin En / Yap Jia Xin