Assignment Title: Federated Learning for Chatbots

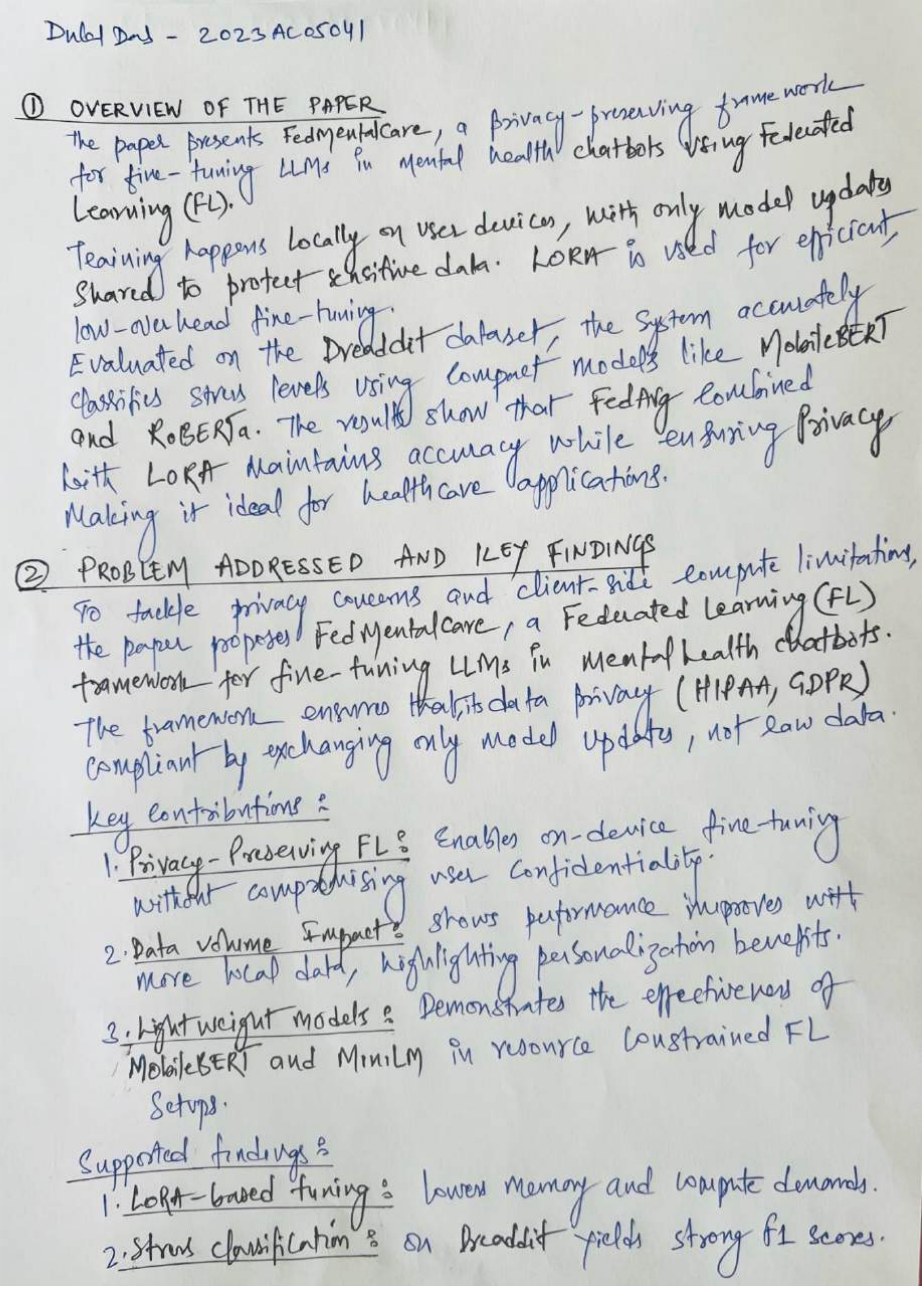
Report on

FedMentalCare: Towards Privacy-Preserving Fine-Tuned LLMs to Analyze Mental Health Status Using Federated Learning Framework

BITS Student ID: 2023AC05041
Name: Dulal Das
Group Number: 43

Title of the Research Paper: *FedMentalCare: Towards Privacy-Preserving Fine-Tuned LLMs to Analyze Mental Health Status Using Federated Learning Framework*

Author: S M Sarwar
Published: March 2025
Paper Link: https://arxiv.org/pdf/2503.05786v2



Dulal Das - 2023AC05041

① OVERVIEW OF THE PAPER

The paper presents FedMentalCare, a privacy-preserving framework for fine-tuning LLMs in mental health chatbots using Federated Learning (FL).

Training happens locally on user devices, with only model updates Shared to protect sensitive data. LORA is used for efficient, low-overhead fine-tuning.

Evaluated on the Dreaddit dataset, the System accurately classifies stress levels using compact models like MobileBERT and RoBERTa. The results show that FedAvg combined with LORA maintains accuracy while ensuring privacy Making it ideal for healthcare applications.

② PROBLEM ADDRESSED AND KEY FINDINGS

To tackle privacy concerns and client-side compute limitations, the paper proposes FedMentalCare, a Federated Learning (FL) framework for fine-tuning LLMs in mental health chatbots.

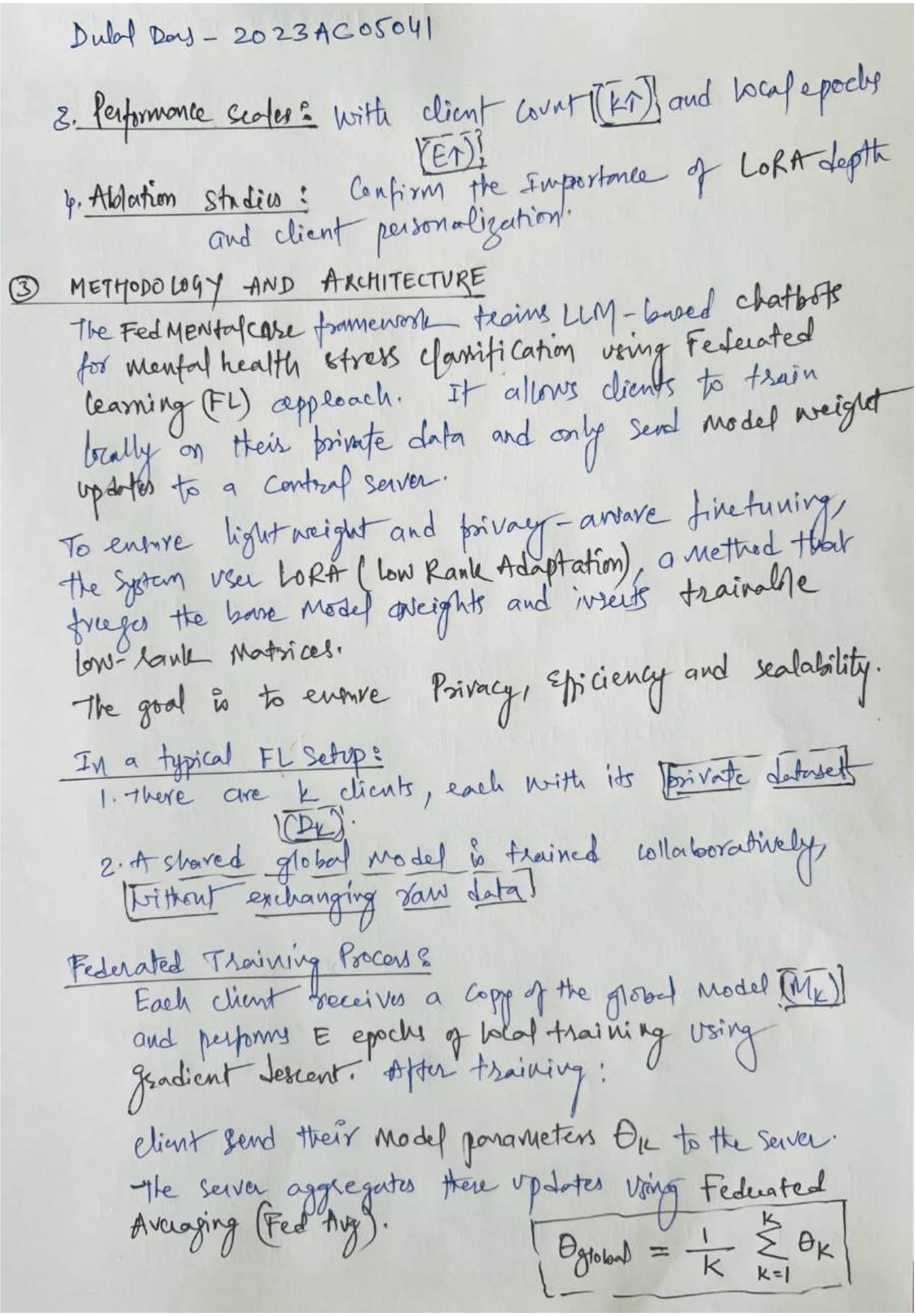
The framework ensures that its data privacy (HIPAA, GDPR) compliant by exchanging only model updates, not raw data.

Key contributions:

1. Privacy-Preserving FL: Enables on-device fine-tuning without compromising user confidentiality.
2. Data volume Impact: Shows performance improves with more local data, highlighting personalization benefits.
3. Lightweight models: Demonstrates the effectiveness of MobileBERT and MiniLM in resource constrained FL Setups.

Supported findings:

1. LoRA-based tuning: lowers memory and compute demands.
2. Stress classification: on Dreaddit yields strong F1 scores.



Dulal Das - 2023AC05041

3. Performance scales: with client count (K↑) and local epochs (E↑).

4. Ablation studies: Confirm the importance of LoRA depth and client personalization.

③ METHODOLOGY AND ARCHITECTURE

The FedMENtalCARE framework trains LLM-based chatbots for mental health stress classification using Federated learning (FL) approach. It allows clients to train locally on their private data and only send model weight updates to a central server.

To ensure lightweight and privacy-aware finetuning, the system uses LoRA (Low Rank Adaptation), a method that freezes the base model weights and inserts trainable low-rank matrices.

The goal is to ensure Privacy, efficiency and scalability.

In a typical FL Setup:

1. there are k clients, each with its private dataset (D_k).
2. A shared global model is trained collaboratively, without exchanging raw data.

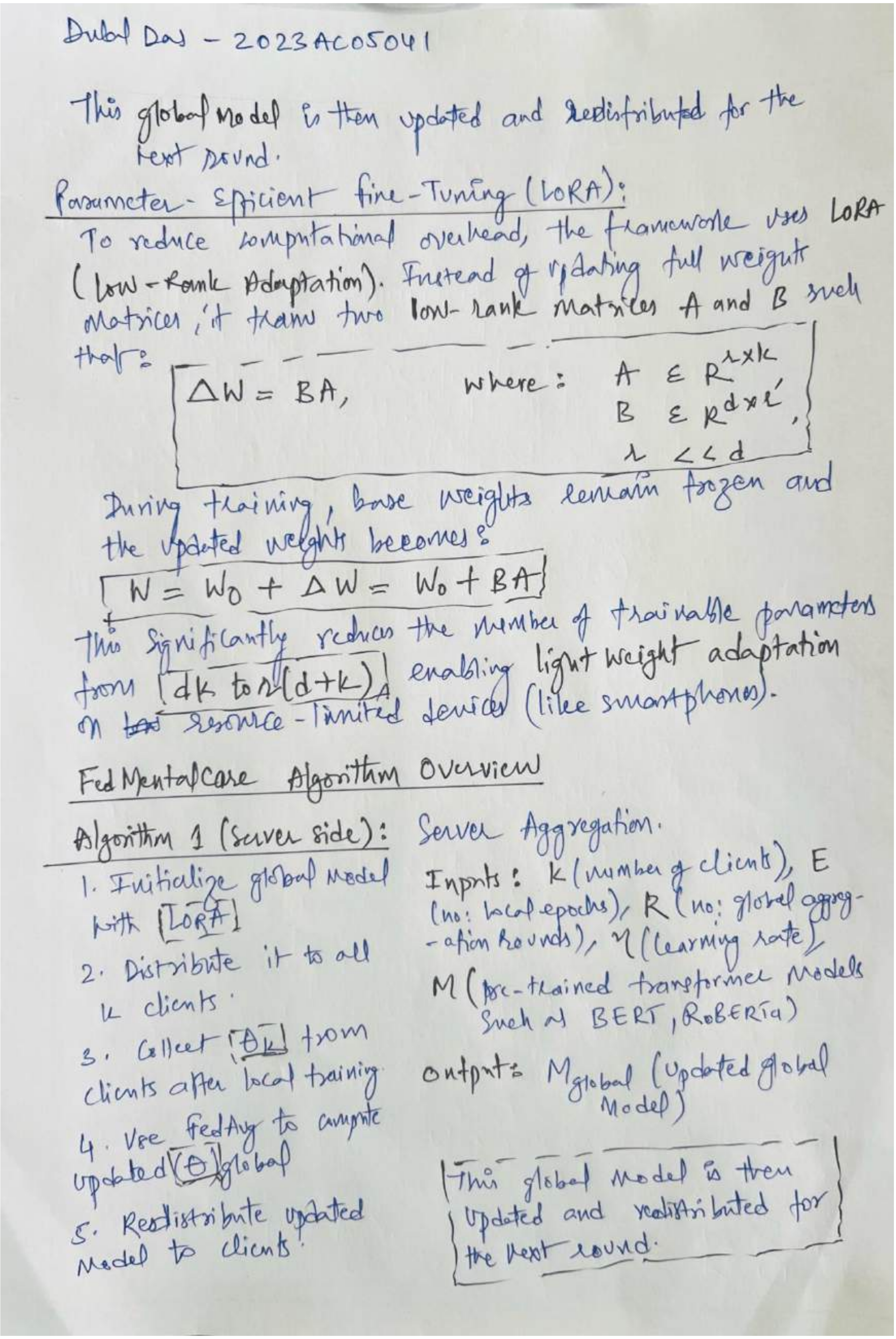
Federated Training Process:

Each client receives a copy of the global model (M_k) and performs E epochs of local training using gradient descent. After training:

client send their model parameters θ_k to the server.

The server aggregates these updates using Federated Averaging (Fed Avg).

$$\boxed{\theta_{global} = \frac{1}{K} \sum_{k=1}^{K} \theta_K}$$



Dulal Das - 2023AC05041

This global model is then updated and redistributed for the next round.

Parameter - Efficient fine-Tuning (LoRA):

To reduce computational overhead, the framework uses LoRA (Low-Rank Adaptation). Instead of updating full weights matrices, it trains two low-rank matrices A and B such that:

$$\boxed{\Delta W = BA, \quad \text{where: } A \in R^{r \times k}, \; B \in R^{d \times r}, \; r << d}$$

During training, base weights remain frozen and the updated weights becomes:

$$\boxed{W = W_0 + \Delta W = W_0 + BA}$$

This significantly reduces the number of trainable parameters from $\boxed{dk \text{ to } r(d+k)}$, enabling light weight adaptation on ~~low~~ resource-limited devices (like smartphones).

FedMentalCare Algorithm Overview

Algorithm 1 (server side): Server Aggregation.

1. Initialize global model with $\boxed{\text{LoRA}}$
2. Distribute it to all k clients.
3. Collect $\boxed{\theta_k}$ from clients after local training.
4. Use FedAvg to compute updated $\boxed{\theta}$ global
5. Redistribute updated model to clients.

Inputs: k (number of clients), E (no: local epochs), R (no: global aggregation rounds), η (learning rate), M (pre-trained transformer models such as BERT, RoBERTa)

output: M_{global} (Updated global Model)

This global model is then Updated and redistributed for the next round.

Dulal Das - 2023AC05041

Algorithm 2 (client-side): Client Training

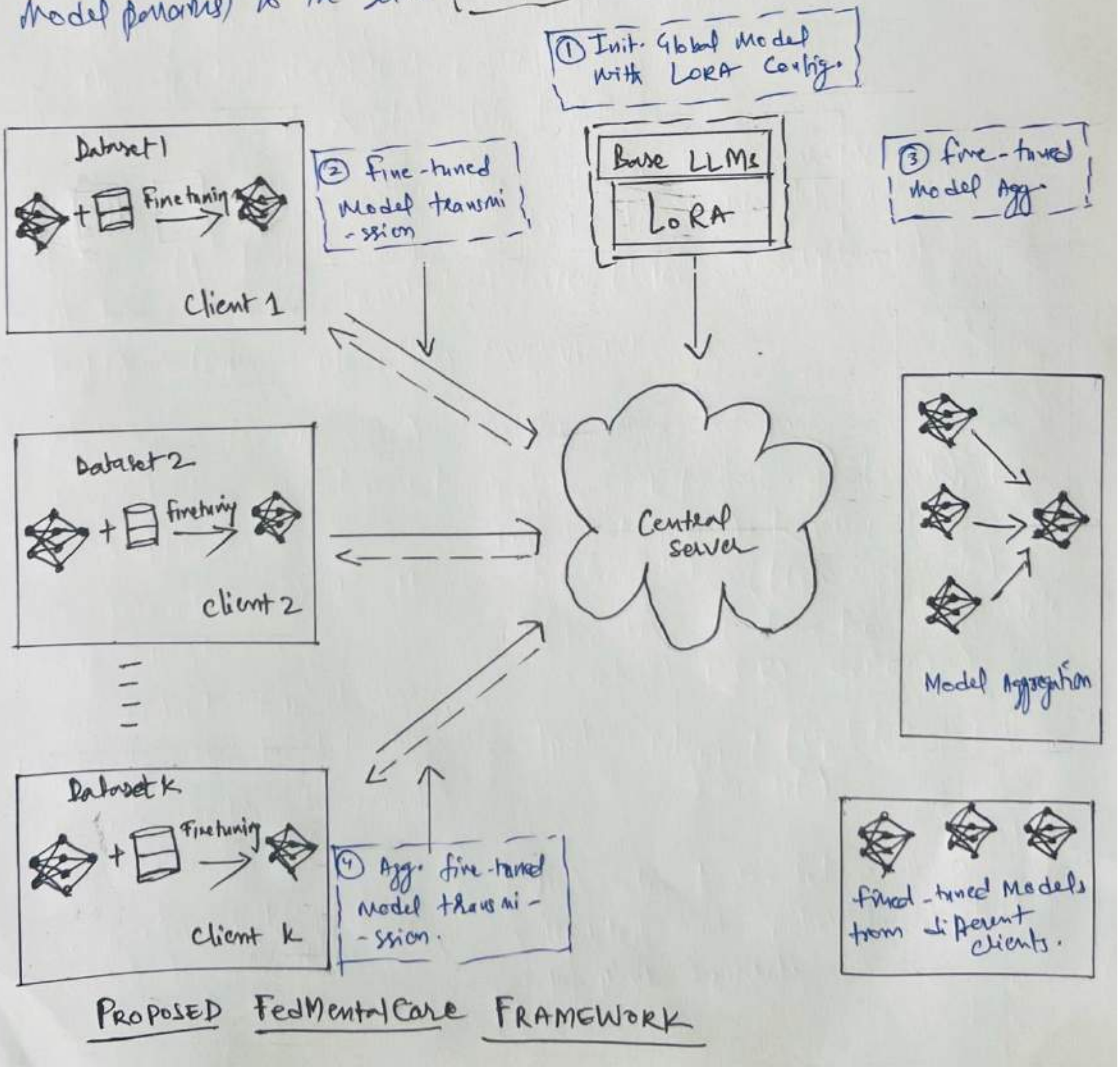
1. Split local data [D_k] into training and validation
2. Tokenize inputs using Model tokenizer.
3. Train for [E] epochs, updating Model weights using gradient descent.
4. Return [Θ_k] (updated model params) to the server

Inputs: D_k (local dataset for each client k), E (number of local training epochs), η (learning rate), T (tokenizer corresponding Models)

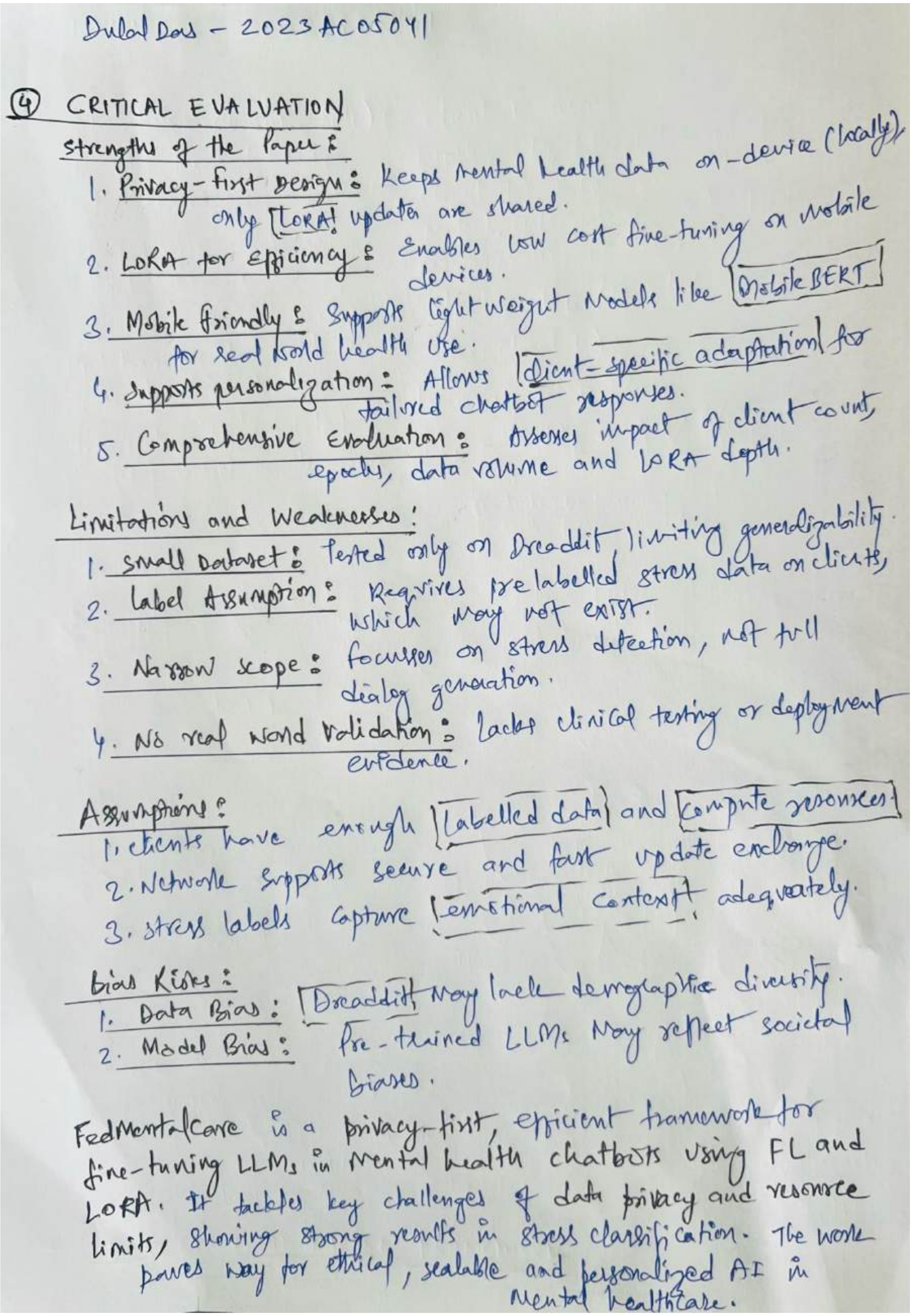
output: Θ_k (Updated model parameters for client k)

Training step 3.

$$M_k \leftarrow M_k - \eta \nabla l(M_k; b)$$


① Init. Global Model with LoRA Config.
Dataset 1
Finetuning
Client 1
② Fine-tuned Model transmi-ssion
Base LLMs
LoRA
③ fine-tuned model Agg.
Dataset 2
finetuning
client 2
Central Server
Model Aggregation
Dataset k
Finetuning
Client k
④ Agg. fine-tuned model transmi-ssion.
fined-tuned Models from different clients.

PROPOSED FedMentalCare FRAMEWORK



Dulal Das - 2023AC05041

④ CRITICAL EVALUATION

Strengths of the Paper:

1. Privacy-first Design: Keeps mental health data on-device (locally), only LoRA updates are shared.
2. LoRA for Efficiency: Enables low cost fine-tuning on mobile devices.
3. Mobile friendly: Supports light weight models like MobileBERT for real world health use.
4. Supports personalization: Allows client-specific adaptation for tailored chatbot responses.
5. Comprehensive Evaluation: Assesses impact of client count, epochs, data volume and LoRA depth.

Limitations and Weaknesses:

1. Small Dataset: Tested only on Dreaddit, limiting generalizability.
2. Label Assumption: Requires prelabelled stress data on clients, which may not exist.
3. Narrow scope: focusses on stress detection, not full dialog generation.
4. No real world validation: Lacks clinical testing or deployment evidence.

Assumptions:

1. clients have enough labelled data and compute resources.
2. Network supports secure and fast update exchange.
3. stress labels capture emotional context adequately.

Bias Risks:

1. Data Bias: Dreaddit may lack demographic diversity.
2. Model Bias: Pre-trained LLMs may reflect societal biases.

FedMentalCare is a privacy-first, efficient framework for fine-tuning LLMs in mental health chatbots using FL and LoRA. It tackles key challenges of data privacy and resource limits, showing strong results in stress classification. The work paves way for ethical, scalable and personalized AI in mental healthcare.