Group 1 Report

Meshari Alatawi
King's College London
meshari.alatawi@kcl.ac.uk

Manvi Agarwal
King's College London
manvi.agarwal@kcl.ac.uk

Denis Baldakov
King's College London
denis.baldkov@kcl.ac.uk

Lara Kanj
King's College London
lara.kanj@kcl.ac.uk

Ethan Davey
King's College London
ethen.davey@kcl.ac.uk

Oluwanifesimi Abolade
King's College London
oluwanifesimi.abolade@kcl.ac.uk

March 26, 2025

1 Abstract

This project entails the development of an AI Diagnosis Assistant, utilising deep learning models and large language models to analyse patient data and suggest potential diagnoses as well as recommending treatments and therapies. This system incorporates analysing text input to extract meaningful symptoms, a ranked diagnosis engine with confidence scores and a recommendation system for next steps. Data collection was done from publicly available medical datasets and these were preprocessed to ensure quality inputs for model training. To enhance diagnostic accuracy, experiments were done with fine-tuning $\text{Bio}_{C}linicalBERTaswellasdevelopingacustomhybriddeeplearningmodel.Userscanbedoctorsorpatientsandthesystemintegr$
$friendlythroughtheinteractiveUIcreatedusingGradio.Keyfeaturesofthisimplementationincludeuserauthentication, chat$
$https://huggingface.co/spaces/Meshari21/AI_Projecthttps://huggingface.co/spaces/Meshari21/AI\_Project$

Data Sourcing and Exploratory Data Analysis (Primary Dataset)

We carefully selected the publicly available **Sentiment Analysis for Mental Health** dataset on Kaggle for its size and diversity. It comprises tens of thousands of real-world mental health statements (chatbot conversations, social media posts, etc.), each clearly labeled, providing a robust foundation for diagnostic predictions.

Preprocessing. We validated the `statement` and `status` columns, removed invalid or missing labels, and cleaned the text by lowercasing, removing URLs, and filtering extraneous characters. We deliberately avoided stop word removal and lemmatization, leveraging BERT's sub-word tokenization for nuanced context. The data was then split into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class distributions. Duplicate statements were removed to reduce overfitting, and random shuffling minimized sampling bias.

Exploratory Data Analysis (EDA). Key findings include:

- **Dataset Size:** Over 33k training samples, with no missing values.
- **Label Distribution:** Stratified sampling yields consistent proportions (e.g., ∼33% Normal, ∼31% Depression).
- **Text Length:** Depression and Suicidal entries often contain longer statements, adding contextual depth.
- **Linguistic Patterns:** TF–IDF and n-gram analyses reveal label-specific phrases (e.g., "kill myself" in Suicidal).
- **Sample Review:** Random checks confirm authenticity and relevance for mental health diagnostics.

Despite these strengths, there were some challenges the EDA underlined that would need careful consideration in model development:

1.1 Secondary Dataset

To supplement the primary dataset, we utilised the Kaggle "Mental Health Diagnosis and Treatment Monitoring Dataset" (500 rows). This dataset provides additional information on diagnoses, treatments, medications and patient outcomes, enhancing the model's ability to generate treatment recommendations based on user inputs.

Data preprocessing was essential to ensure data consistency and accuracy for model training. The first step in this process was data cleaning, where duplicates, missing values, invalid numerical values and incorrect categorical values were identified and either removed or corrected to ensure the dataset was in a usable format for analysis.

Another key aspect of data preprocessing was the mapping of diagnoses between two datasets, which have different diagnostic labels. To maintain consistency across datasets, specific mappings were applied.

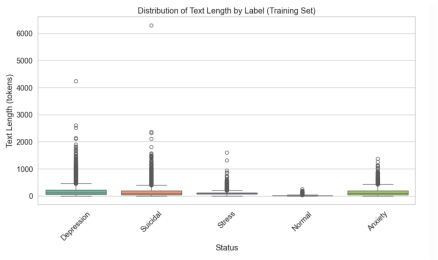

Figure 1: Some entries exceed BERT's 512-token limit. (There are also extremely short texts of one or two words)

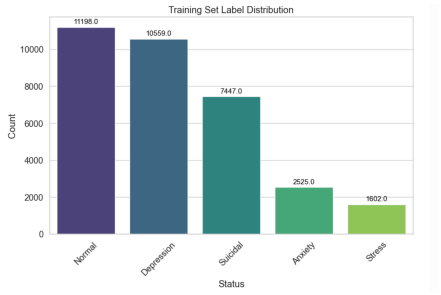

Figure 2: Imbalance due to minority classes (Anxiety, Stress) under 10%.

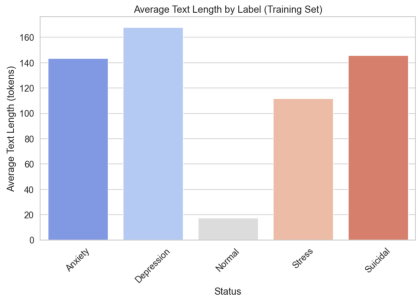

Figure 3: Normal statements are typically shorter, posing a length bias risk.

Bipolar Disorder was mapped to **Bipolar**; **Panic Disorder** and **Generalised Anxiety Disorder** were both mapped to **Anxiety**; and **Major Depressive Disorder** was mapped to **Suicidal** if the "Symptom Severity" was greater than or equal to 8, or to **Depression** if the "Symptom Severity" was less than or equal to 7. The **Normal** diagnosis did not require any mapping, as it resulted in no treatment recommendation.

The cleaned dataset was divided into training (70%), testing (15%) and validation (15%) sets, maintaining the same ratio as the first dataset to ensure consistency in model training.

1.1.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted on the cleaned datasets to uncover key patterns in the data. The analysis began with a distribution study of key features, such as **diagnosis, gender, and age. Anxiety** was found to be the most common diagnosis across the training, testing, and validation datasets. The gender distribution was nearly equal, although there was a slightly higher representation of male patients. In terms of age, the majority of patients were between 25-39 years old, classifying them as adults.

The correlation between diagnoses and medications was also examined. The most commonly prescribed medications were specific to each diagnosis: **SSRIs** were most commonly prescribed for **Anxiety**, **Anxiolytics** for **Bipolar Disorder**, **Benzodiazepines** for **Depression**, and both **Benzodiazepines** and **Anxiolytics** for patients diagnosed as **Suicidal**.

The effectiveness of treatments was evaluated by assigning numerical values to patient outcomes. This allowed for the calculation of average effectiveness scores for various medications and therapies. The results showed that **Mood Stabilizers** were the most effective medication, while **Antidepressants** were found to be the least effective. In terms of therapy, **Cognitive Behavioural Therapy (CBT)** was identified as the most effective, while **Mindfulness-Based Therapy** was the least effective.

Finally, a correlation analysis was conducted to assess the relationship between treatment progress and

numerical features such as symptom severity and therapy mode scores. The analysis revealed that **Sleep Quality** had the strongest positive correlation with **Treatment Progress**, suggesting its significant impact on patient recovery.

2 Model Development

Diagnosis Prediction and Explanation

System Architecture

We developed a dual-model architecture for mental health diagnosis classification and explanation generation. We use **BioClinicalBERT** for the diagnosis classification task, which has been pretrained on all notes from **MIMIC III**, a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, USA.

We then use fine-tuned OpenAI GPT models to provide explanations on top of these predicted diagnoses, helpfully connecting the patient's input with our model output. The system achieves high diagnostic accuracy across five mental health categories while providing clinically relevant explanations for practitioners. We obtained **over 85% top-1 accuracy** on the diagnosis prediction task, jumping to **over 97% for top-3 accuracy**.

The decoupled design of our system architecture enables specialized optimization of each component while maintaining a coherent end-to-end system with the workflow: **Patient Statement** $\rightarrow$ **BERT Classifier** $\rightarrow$ **Diagnostic Prediction** $\rightarrow$ **Clinical Explanation Generator**, **Medication and Treatment Predictions**.

BERT Classification

After extensive comparison between models, **BioClinicalBERT** was selected for its pre-training on clinical text and superior performance on healthcare NLP tasks. We fine-tuned the model provided at `huggingface.co/emilyalsentz` on the **Sentiment Analysis for Mental Health** dataset outlined above.

The training regime followed a multi-phase approach with initial transfer learning, intermediate evaluation, and final fine-tuning, and optimization through **Optuna** identified optimal hyperparameters. All training experiments were recorded and tracked using **Weights and Biases**. Comprehensive experiment tracking included performance evaluation, model artifact storage, hyperparameter optimization, and hardware utilization monitoring.

GPT Explanation Generation

We optimized **GPT-4o-mini** through prompt engineering and fine-tuning. The system prompt positions the model as "an expert mental health assistant" serving as "a supplementary tool to the doctor or medical professional," focusing on concise, evidence-based explanations using the relevant medical terminology. Explanation quality was subjectively assessed by us during development.

Medication and Treatment Prediction

We developed a hybrid deep learning model utilizing **BioClinicalBERT** to predict both **medication** and **therapy** recommendations based on the previously predicted mental health diagnosis. Alongside the diagnosis, **age** and **gender** were incorporated as additional features to enhance the model's predictive performance.

The architecture consists of:

- A BERT-based encoder to process the diagnosis text.
- An MLP (Multi-Layer Perceptron) layer to process age as a numerical feature.
- An embedding layer to process gender as a categorical feature.

test_predictions.csv ×

1 to 10 of 75 entries Filter

Diagnosis	True Medication	Predicted Medication	True Therapy	Predicted Therapy
Anxiety	Anxiolytics	SSRIs	Cognitive Behavioral Therapy	Dialectical Behavioral Therapy
Suicidal	Benzodiazepines	Antipsychotics	Mindfulness-Based Therapy	Interpersonal Therapy
Bipolar	Antidepressants	Antipsychotics	Interpersonal Therapy	Mindfulness-Based Therapy
Anxiety	SSRIs	SSRIs	Dialectical Behavioral Therapy	Dialectical Behavioral Therapy
Suicidal	Benzodiazepines	Antipsychotics	Interpersonal Therapy	Interpersonal Therapy
Anxiety	Mood Stabilizers	SSRIs	Interpersonal Therapy	Dialectical Behavioral Therapy
Anxiety	SSRIs	SSRIs	Cognitive Behavioral Therapy	Dialectical Behavioral Therapy
Anxiety	Antidepressants	SSRIs	Interpersonal Therapy	Dialectical Behavioral Therapy
Depression	Mood Stabilizers	Benzodiazepines	Interpersonal Therapy	Cognitive Behavioral Therapy
Bipolar	Mood Stabilizers	Antipsychotics	Dialectical Behavioral Therapy	Mindfulness-Based Therapy

Show 10 per page 1 2 3 4 5 6 7 8

Figure 2: Example Medication Predictions

These components are concatenated and passed through fully connected layers, from which the model outputs predictions for both medication and therapy types. This hybrid setup allows for extensibility, enabling the future integration of additional feature types with minimal architectural changes.

To address the **class imbalance** issue flagged during data preparation, we applied `compute_class_weight` from Scikit-learn to generate class weights and used **weighted cross-entropy loss**. This approach ensured that underrepresented classes retained a fair influence during training, reducing bias toward more common medication categories. We also standardized age using **z-normalization** to mitigate scale differences.

Hyperparameter tuning included:

- Experimenting with optimizers and learning rates — the best performance was achieved using the **AdamW** optimizer with a learning rate of `1e-5`.
- Varying hidden layer sizes and dropout rates to prevent overfitting while maintaining model complexity.
- Adjusting batch size based on memory constraints and training time.

Model Evaluation and Challenges

Despite extensive tuning, the model's performance was limited by data constraints. The final training dataset contained only **350 rows** after cleaning, which proved insufficient for a deep learning model to learn meaningful patterns. In addition, the challenge of **one-to-many mappings** (i.e., a single diagnosis mapping to multiple valid medications and therapies) complicated the learning process.

The class imbalance persisted, and even with weighted loss, the model struggled to generalize. However, the hybrid architecture provided better representation across diverse feature types and showed potential for future scalability.

Model evaluation metrics included:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**

We also analyzed **confidence scores** and tested the model on an unseen test set obtained via an internal data split. These results highlighted both the model's limitations and its potential with a larger and more diverse dataset. Please refer to Figure 2 which contains examples of the predicted medications and therapies.

3 User Interface Design and Deployment

Introduction

The user interface (UI) was developed using Gradio, a framework particularly helpful for AI and ML projects due to its simplicity and ease of integration. The system includes several key features, such as user authentication, chat history tracking, appointment booking, and automated PDF report generation, all designed to improve the overall user experience and functionality of the platform. Additionally, we added a basic CSS stylesheet to enhance the visual design and layout, allowing for more customized styling beyond Gradio's default appearance. The final system was deployed on Hugging Face Spaces, where it became publicly accessible through a shared link, with environment variables like the OpenAI API key securely managed using the platform's secrets feature.

Features

- **Login/Register:** The entry point of the application, allowing users to log in using existing credentials or register for a new account. Usernames must be unique, passwords must be longer than 8 characters, and emails must be in a valid format. Account information is securely stored in a database, with passwords encrypted to ensure safe user sessions.
- **Main Page:** Users can input patient information—name, age, gender, and symptoms—on the left-hand side, and receive diagnosis results, treatment recommendations, and more on the right-hand side. A feature is also provided to generate a report in PDF format.
- **Chat History:** Displays previous interactions with the chatbot, along with timestamps, allowing users to track patient progress or review past sessions.
- **Appointment Booking:** Enables users to schedule appointments by specifying the date, time, and reason for the visit.
- **Profile Selection:** By clicking the profile button, users can view their account information and access the logout option to return to the login screen.

UI Design Stages and Final Result

The initial version of the user interface was quite basic and lacked elements that would make it user-friendly or visually appealing. However, as the project progressed, we added improvements to the design — as shown in Figure 6. This intermediate version had a decent design but we recognized that it still needed more work. In our final design phase, following suggestions provided by our advisor during the last meeting, we implemented key changes that significantly improved the overall look and usability of the interface.

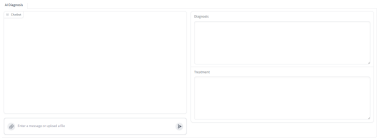

Figure 5: Initial User Interface

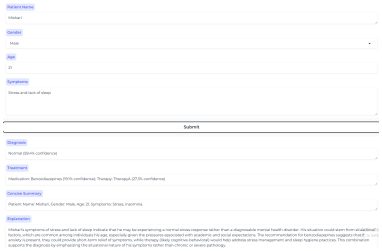

Figure 6: Intermediate User Interface

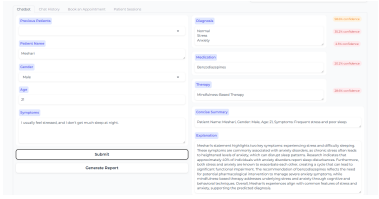

Figure 7: Final User Interface

Project Management

Planning. The team held two fixed weekly meetings to align on the business goal, clarify deliverables, share progress, and address any challenges collaboratively. Work was divided based on each member's strengths,

with roles grouped into sub-teams (e.g., data, model development). We used Jira to efficiently manage tasks, track progress, and visualize the workflow, which helped maintain a clear structure throughout the project.

Monitoring. Weekly goals were set alongside the long-term objective to keep the project on track and ensure efficient progress. Team calls were held whenever blockers arose, enabling us to either solve issues promptly or adjust targets as needed. Cross-team communication was consistent, with members frequently sharing feedback and updates to ensure mutual understanding and smooth integration of components. Code was regularly pushed to GitHub, with clear `README` documentation per branch, and features were tested across devices and setups to ensure robustness.

Risk Assessment. From the outset, we identified several key risks, including data quality, dataset sufficiency, and handling edge cases like non-diagnostic inputs. Extensive research and exploratory data analysis helped validate our data choices and understand potential weaknesses. We also built logic to detect and manage incorrect or intentionally misleading inputs, and privacy safeguards were implemented to protect user history and sensitive content.

Challenges and Solutions. Integrating multiple models into a seamless system was complex, but this was resolved by first ensuring each component worked independently, then collaboratively testing their interoperability. When data issues were uncovered during EDA, the data team led a dedicated session to present findings, followed by group research and strategic discussion. The chosen solutions were then implemented by the model development team, leading to effective resolution aligned with the project's goals.

Additionally, due to the limitations of GitHub Enterprise, we were unable to use automated testing for the project. To overcome this challenge, we implemented an extensive manual testing process to thoroughly evaluate the functionality of the code. Our unit testing covered all critical aspects of the project, from data pre-processing to model development, and the coverage statistics demonstrated strong results, ensuring comprehensive test coverage.