Submitted to

**The University of Roehampton**

In partial fulfilment of the requirements

for the degree of

**Master of Science**

**in**

**Data Science**

By

**Beneyaz Samium Amin Rakib**

**Digital Twins for Healthcare: A Secondary Data-Driven Approach to Mitigating Mental Stress**

Declaration

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Beneyaz Samium Amin Rakib

14/08/2025

BSA RAKIB

Acknowledgements:

For the successful completion of this project, I would like to thank my supervisor Sameena Naaz for her guidance all the way through. I would also like to thank the university of Roehampton and the professors who involved in the workshops given to us throughout the semesters.

Abstract:

It is a smart mental health helper that identifies various forms of stress based on user input and provides the user with compassionate responses. It categorizes stress as academic stress, relationship stress, anxiety among others. Lot of people cannot share emotional strain or do not obtain accessible support in time. The automation of stress detection at an early stage and the ability to deliver compassionate reactions can make the users feel listened-to and advice, even in cases when human interaction is not readily accessible in a timely manner. Sentence-BERT model is used to translate an initial text to a collection of semantic embeddings which are, in turn, classified by an XGBoost model which was trained on stress data labels. The backend uses the text of the predicted type of stress to create a personalized prompt and asks Groq language model (based on LLaMA) to give a compassionate reply. This whole system was constructed using FastAPI and will have the capability to communicate in real-time through a frontend interface

Table of Contents

[Declaration ii](#_Toc205897712)

[Acknowledgements iii](#_Toc205897713)

[Abstract: iv](#_Toc205897714)

[Table of Contents v](#_Toc205897715)

[List of Figures ix](#_Toc205897716)

[List of Tables x](#_Toc205897717)

[CHAPTER 1 - Introduction 1](#_Toc205897718)

[1.1 Problem Description, Context and Motivation 1](#_Toc205897719)

[1.2 Objectives 1](#_Toc205897720)

[1.3 Methodology 2](#_Toc205897721)

[1.3.1 Design: 2](#_Toc205897722)

[1.3.2 Trials and Testing: 2](#_Toc205897723)

[1.4 Project Management: 2](#_Toc205897724)

[1.4.1 Technologies and Processes: 3](#_Toc205897725)

[1.4.2 Legal, Social, Ethical and Professional Considerations 3](#_Toc205897726)

[1.5 Background 4](#_Toc205897727)

[1.6 Structure of Report 4](#_Toc205897728)

[Chapter 2 Literature – Technology Review 6](#_Toc205897729)

[2.1 Literature Review 6](#_Toc205897730)

[2.1.1 Problem Description and Context 6](#_Toc205897731)

[2.1.2 Summary and Discussion of Key Literature 6](#_Toc205897732)

[2.1.3 Relevance to Problem Statement 7](#_Toc205897733)

[2.1.4 Critical Evaluation 7](#_Toc205897734)

[2.1.5 Conclusions and Guidance for Methodology 7](#_Toc205897735)

[2.2 Technology Review 7](#_Toc205897736)

[2.3 Summary 9](#_Toc205897737)

[Chapter 3 Implementation 10](#_Toc205897738)

[3.1 Implementation 10](#_Toc205897739)

[3.2 System Design & Implementation: 10](#_Toc205897740)

[3.3 Sprint Breakdown 11](#_Toc205897741)

[3.4 Dataset Acquisition and Cleaning 12](#_Toc205897742)

[3.5 EDA 13](#_Toc205897743)

[3.6 Feature Engineering 18](#_Toc205897744)

[3.6.1 TF – IDF Vectorisation 18](#_Toc205897745)

[3.6.2 Sentence BERT Embedding 19](#_Toc205897746)

[3.7 Baseline Models 19](#_Toc205897747)

[3.7.1 Random Forest 19](#_Toc205897748)

[3.8 Advanced Model 20](#_Toc205897749)

[3.8.1 SBERT + XGBoost 20](#_Toc205897750)

[3.9 Frontend Implementation (React.js & Tailwind CSS) 20](#_Toc205897751)

[3.10 Backend Implementation (FastAPI) 21](#_Toc205897752)

[3.11 Conversational Response Using LLM (Groq API) 23](#_Toc205897753)

[3.12 Full System Workflow 24](#_Toc205897754)

[3.13 Difficulties and Answers 25](#_Toc205897755)

[3.14 Ethical Considerations and Limitations 26](#_Toc205897756)

[3.15 Future Enhancements 27](#_Toc205897757)

[Chapter 4 Evaluation and Results 28](#_Toc205897758)

[4.1 Related work: 28](#_Toc205897759)

[4.2 Evaluation Methodology 29](#_Toc205897760)

[4.2.1 Model Performance Testing 29](#_Toc205897761)

[4.2.2 System Performance Testing 30](#_Toc205897762)

[4.3 Quantitative Results 30](#_Toc205897763)

[4.3.1 Stress Classification Accuracy 30](#_Toc205897764)

[4.3.2 Confusion Matrix 32](#_Toc205897765)

[4.3.3 API & Backend Response Time 33](#_Toc205897766)

[4.3.4 Frontend Load & Interaction Metrics 33](#_Toc205897767)

[4.4 Qualitative Feedback: 34](#_Toc205897768)

[4.4.1 Results 34](#_Toc205897769)

[4.5 Evaluation of LLM-Based Responses 34](#_Toc205897770)

[4.5.1 Criteria for Evaluation 34](#_Toc205897771)

[4.6 Results: 34](#_Toc205897772)

[4.7 Error Analysis 35](#_Toc205897773)

[4.7.1 Classification Errors 35](#_Toc205897774)

[4.7.2 LLM Errors 35](#_Toc205897775)

[4.8 Usability and UX Testing 36](#_Toc205897776)

[4.9 Ethical Testing and Feedback 36](#_Toc205897777)

[4.9.1 Privacy Simulation 36](#_Toc205897778)

[4.9.2 Ethical Considerations 36](#_Toc205897779)

[4.10 Conclusion 37](#_Toc205897780)

[Chapter 5 Conclusion 38](#_Toc205897781)

[5.1 Summary of Objectives and Achievements 38](#_Toc205897782)

[5.2 Research Contributions 38](#_Toc205897783)

[5.2.1 Technological Innovation 38](#_Toc205897784)

[5.2.2 Psychological Utility 39](#_Toc205897785)

[5.3 Critical Reflections 39](#_Toc205897786)

[5.3.1 Uncertainty with regards to Stress Classification 39](#_Toc205897787)

[5.4 Cultural and Linguistic constraints 40](#_Toc205897788)

[5.5 Broader Implications 40](#_Toc205897789)

[5.6 Future Work 41](#_Toc205897790)

[5.6.1 Voice Integration 41](#_Toc205897791)

[5.6.2 Emotion-Aware Chat 41](#_Toc205897792)

[5.6.3 Multilingual Support 41](#_Toc205897793)

[5.6.4 Offline Capabilities 41](#_Toc205897794)

[5.6.5 Integration with Medical Platforms 41](#_Toc205897795)

[5.6.6 Progress Tracking 41](#_Toc205897796)

[5.7 Final Thoughts 42](#_Toc205897797)

[References 43](#_Toc205897798)

[Appendices I](#_Toc205897799)

[Appendix A: Project Proposal II](#_Toc205897800)

[Appendix B: Project Management III](#_Toc205897801)

[Appendix C: Artefact/Dataset IV](#_Toc205897802)

[Appendix D: Screencast V](#_Toc205897803)

List of Figures

[Figure 1:Gantt Chart 3](#_Toc205897303)

[Figure 2:Methodology 11](#_Toc205897304)

[Figure 3:Label Distribution 14](#_Toc205897305)

[Figure 4:Text Length Distribution 15](#_Toc205897306)

[Figure 5:Text Length by Label 16](#_Toc205897307)

[Figure 6:Most Frequent Words 17](#_Toc205897308)

[Figure 7:Character Count Distribution 18](#_Toc205897309)

[Figure 8:Text Vectorization code 19](#_Toc205897310)

[Figure 9:Sentence Bert Embedding 19](#_Toc205897311)

[Figure 10:Model Classification Report 32](#_Toc205897312)

[Figure 11:Model AUC-ROC curves 32](#_Toc205897313)

[Figure 12:Confusion Matrix of Xgboost Model 33](#_Toc205897314)

List of Tables

[Table 1:Sprint Breakdown 12](#_Toc205329141)

[Table 2:Text PreProcessing Workflow 13](#_Toc205329142)

[Table 3: System design and stack 20](#_Toc205329143)

[Table 4:Stress Detection and Response Workflow 21](#_Toc205329144)

[Table 5: Comparative Benchmarks 25](#_Toc205329145)

# CHAPTER 1 - Introduction

## Problem Description, Context and Motivation

During recent years, mental health concerns and, especially, stress, anxiety, and depression have acquired the threat to the population health. The ever-accelerating rhythm of life and the digital overload psychosis, in combination with social-economic insecurity, have played a role in creating many citizens who are chronically stressed (World Health Organization, 2021). This is a major concern especially to people who might not get timely psychological support or those who fail to apply themselves to receive help because of the social stigma (Guntuku et al., 2017).

The issue that is being solved by this project is the fact that mental stress is promptly and efficiently diagnosed using non-invasive methods, namely by examining the text of the conversation produced by a user. Both victims of this situation are those users of social media, particularly those perpetrators of mental health-related messages, e.g., in the one of the mental health forums, Reddit. The issue arises when people are in online settings and communicate openly or implicitly their state of mind by using some textual communication.

This issue is significant because early interventions can be used in its solution. The recognition of stress by routine communication can allow the creation of responsive support systems, which report on mental welfare constantly and passively. These systems can provide interventions in time or escalating them before the condition exacerbates which means improved mental health and less of a burden on the healthcare system. By using natural language processing (NLP) and machine learning (ML), there are operational solutions, which can be scaled and privacy-aware to be both viable technologically and have social good (Torous et al., 2016).

## Objectives

The goals this project are stated below:

* To pre-process and analyse a labelled Kaggle dataset based on conversational text labelled with stress.
* To train model and implement machine learning, both classic classifier models and transformer-based models like BERT are used, to classify the presence of stress in text.
* Constructing a simple chat interface or program that acts as a digital twin capable of performing real-time prediction of the stress based on user behaviour.
* To assess the model based on generic measures like accuracy, precision, recall, F1-score, and the AUC.

## Methodology

The project is a data driven one which incorporates machine learning and natural language processing to identify stress. The methodology will be correspondingly arranged:

### Design:

The design commences with preprocessing of data, the cleaning of data, tokenisation, and representation of the data in a numerical-vector technique and at this point TF-IDF and BERT embeddings have been relied upon. A labelled dataset, where variables to be predicted include anxiety, depression, stress, suicidal, and normal, is provided to train and validate classification models to detect mental health.

There are two types of applied models:

* Embedding based ML model: XGBoost stress classification based on BERT embedding.
* Rule based sentiment classification: TextBlob is utilized to evaluate the sentiment of input by the user.

The choice of models is based on previous studies that demonstrate excellent outcomes of transformer models of emotion and stress detection in text (Rastogi et al., 2022).

### Trials and Testing:

Cross-validation is used to test models and have it tested on unseen data to make it generalizable. Evaluation measures are accuracy, precision, recall, F1-score, ROC-AUC. The confusion matrix of the performance is employed.

## Project Management:

Agile approach is embraced along with kanban to organize work. The project will be broken down into weekly sprints, behavioural preparations and model training, interface definition and testing between them will be established.

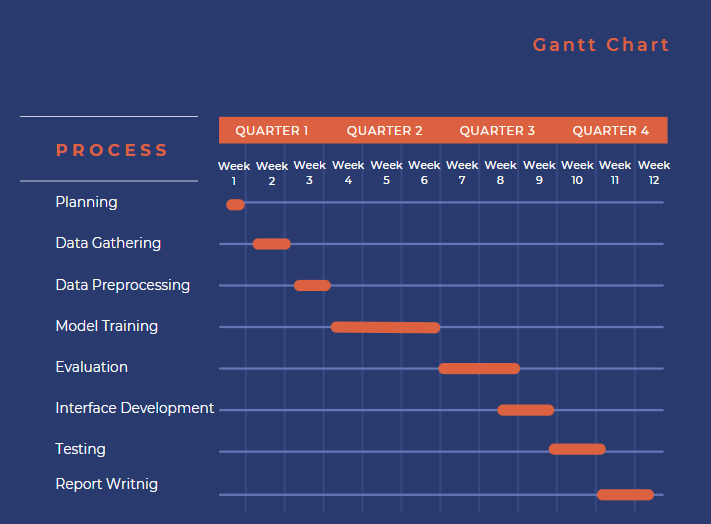


Figure 1:Gantt Chart

### Technologies and Processes:

Language of Programming: Python & React

Scikit-learn, TensorFlow, Transformers (HuggingFace), Pandas, NLTK, Matplotlib

Tools: Google Colab, FastAPI, GitHub.

BERT is chosen due to its contextual interpretation of language as well as the success reported about the performance of similar tasks including sentiment and emotional recognition (Rastogi et al., 2022).

### Legal, Social, Ethical and Professional Considerations

The data is obtained through Kaggle where the information is anonymized and probably on the open sites, such as Reddit, and is GDPR compliant. The system is not diagnosing anyone; it aims at creating awareness which is accompanied by disclaimers. Bias is mitigated with fairness checks and explainability, and transparency, and reproducibility issues, and privacy concerns cannot be overlooked during development (Bruynseels et al., 2018).

## Background

Mental health is a very burning question that concerns citizens of all demographics. As awareness increases, there is a trend towards the early assessment and treatment of mental problems by relying on digital tools. Digital twins have already appeared in healthcare, and there have been positive attempts to model physiological conditions with it; this project is a continuation of such developments but with psychological states (Bruynseels et al., 2018).

The recent studies prove that the accuracy of stress analysis based on language. Through a study by Guntuku et al. (2018), it was revealed that language in social media might forecast stress and mental health consequences. Equally, Reece et al. (2017) described some Instagram posts patterns that corresponded to depressive symptoms. The ML, and especially NLP technologies, including transformer models such as BERT, now allow extracting the emotional hints with a high degree of accuracy, in regard of text (Torous et al., 2016).

In this work, the dataset is based on Kaggle and consists of three features: an ID column, a text column containing user-written text, and a status column that identifies the mental health category. The status options are anxiety, suicidal, normal, depression, and stress. Such kind of data is very conducive to developing digital mental health applications like chatbots and online emotional support systems.

## Structure of Report

The report is written in the following way:

Chapter 2: Literature Review

Reviews the existing studies on stress recognition based on NLP and machine learning, digital twin in health.

Chapter 3: Implementation

This chapter explains the data preparation and preprocessing chain, machine learning algorithms and the development process of the system and interface integration.

Chapter 4: Results and Evaluation

This section provides experimental results, performance metrics, analysis of the errors and a comparison of the models.

Chapter 5: Conclusion

Outlines important findings, contributions, limitations and makes a recommendation of improvement in the future.

# Literature – Technology Review

## Literature Review

### Problem Description and Context

Mental stress affects over 264 million people globally with anxiety disorders (Bhattacharya, Ghosh, & Ghosh, 2021). Currently there is no option for real-time, personalized detection and monitoring of mental stress. Current tools for measuring stress, such as the Perceived Stress Scale (PSS) and Depression Anxiety Stress Scales (DASS), rely on self-reported questionnaires that may exhibit bias due to the use of memory and subjective rating processes, and they cannot be used to be continuously assessed. Digital twin (DT) technology leverages real-time sensor and clinical data to create real-time virtual patient models and can predict and/or assist in the management of a specific aspect of its real-world patient counterpart (Can, Arnrich, & Ersoy, 2019). Digital twins are known for their uses in physical health and mental health has not been fully explored using digital twin technology; however, their use for mental stress management is a major gap in this area, requiring research to assess feasibility.

### Summary and Discussion of Key Literature

DTs leverage real-time data to model physiological processes and facilitate precision medicine. Corral-Acero et al. (2020) explained DTs in cardiology where physiologic data were integrated into the simulation of surgical outcomes, and Bruynseels et al. (2018) described DTTs for oncology, where physiological data was integrated into models of tumor growth .In mental health, while we have some measures for assessing stress (e.g. PSS, DASS), general bias limits their utility. Physiological markers of stress (HRV, GSR, cortisol) provide objective data, but they are rarely incorporated into predictive models [6]. Gjoreski et al. (2017) used a variety of sensor data (HRV, GSR, accelerometry) to detect stress, but did not model via DTs. Can et al. (2019) were able to associate stress to activity measured via smartphone but did not employ DT frameworks [8]. Moreover, Bhattacharya et al. (2021) proposed a DT approach to mental health by integrating behavioral and physiological data, but it remains to be validated. Modeling subjective mental states generally present challenges to DT, and digital supply chains suffer from data heterogeneity (Torous et al., 2016), as well as ethical issues related to privacy under GDPR, and HIPAA.

### Relevance to Problem Statement

Literature supports the issue of inadequate tools for managing stress. The success of DTs in the realm of physical health suggests the potential for continuous, objective stress monitoring and could address the disadvantages of subjective approaches (Jin et al., 2021). Similarly, the sensor-based studies (Rastogi et al., 2022) conducted for this research align with the aim of the study to evaluate the feasibility of the use of DTs for mental stress, while also addressing the technical and ethical gaps.

### Critical Evaluation

DTs are effective in relation to physical wellness but weak in the subjective mental states (Shaban-Nejad, Michalowski, & Buckeridge, 2018). Sensor research shows promise, but lacks DTs (Bruynseels, Santoni de Sio, & van den Hoven, 2018). Bhattacharya et al.'s framework is novel, not proven. Privacy and bias pose problems, and a lack of standards exist [3]; hence the need for solid frameworks and ethical guidelines.

### Conclusions and Guidance for Methodology

The literature supports the value of DTs for stress management but highlights limitations around modeling, data integration and ethics. This study's qualitative secondary data analysis will involve a thematic analysis of reports and case studies to model an DT, focusing on stress detection, interoperability and ethical compliance as appropriate to the study's objectives.

## Technology Review

The digital twin (DT) technology, which emerged for industrial applications in the early 1990s, has safely been adapted to the healthcare domain, especially given the rapid developments in wearable sensing, big data analytics, and cloud computing. A DT in the healthcare domain represents patient and/or healthcare system as a dynamically changing, virtual Model that integrates patient data on a real-time basis via patient-centred data sources including wearable devices, electronic health records, and environmental sensor data (Tao et al., 2018). Developed applications of DTs include the Dassault Systèmes heart DT which simulates surgical scenarios to support better accuracy in cardiology (Corral-Acero et al., 2020). These applications gather a complex data set of imaging, laboratory results, and sensor based data and create high-fidelity Models, enabling predictive analytics and patient-specific interventions.

In the case of DTs that deal with managing mental stress, the aim will be to ensure the DT captures multiple data streams from a variety of sources, including physiological (e.g., HRV, GSR), behavioral (e.g., sleep patterns, activity), and environmental (e.g., noise, light) data sources. These data sources will mainly come from wearable devices, e.g., smartwatches and biosensors like Fitbit or Empatica E4, which are good for continuous data collection for real-time monitoring (Gjoreski et al., 2017), as well as smartphone-based sensing of mobile behavior and apps (Can et al., 2019). Cloud computing platforms, such as AWS or Azure, support the storage and analysis of large data sets, and machine learning (ML) algorithms produce predictive models that could be used in regards to stress, for example, a DT may help by analyzing HRV data trends alongside sleep disturbance to produce stress episode predictions and/or recommendations on intervention, such as mindfulness exercises, or article recommendations for professional consultation, or provide a direct TCP to a resource for professional consultation.

Nevertheless, designing digital twins for mental health has many technical challenges. Data heterogeneity requires flexible interoperability frameworks to combine data from variety of original sources with diverse formats and sampling rates (for example, Jin et al., 2021). Affective computing, in which emotional states can be derived from biometric indicators is also nascent, characterizing a significant limitation in predicting mental states accurately (Calvo & D’Mello, 2010). The use of machine learning algorithms to design digital twins also entails staying conscious of algorithmic bias and ensuring explainability so that outcomes can be equitable and interpretable (Rajkomar et al., 2019). Lastly, there are cybersecurity aspects involved, as mental health data requires the same phenomenally comprehensive protection mechanisms that organizations must follow under the General Data Protection Regulation (GDPR) and processes akin to a Health Information Portability.

Accountability Act (HIPAA), including encryption and de-identification of individuals (Voigt and Von dem Bussche, 2017).

New technologies provide solutions. For example, edge computing has advanced to allow data to be processed in-wearable in real time (a big advantage), as it lowers latency problems and potential privacy concerns. Blockchain technology could also provide data security with transparent and verified hacking-proof records on remote servers. Overall, standard protocols to develop digital twins (DT)s, as proposed by Tao et al. (2018), could facilitate interoperability among DTs and clinical adoption into practice.

## Summary

Literature and technology reviews show that healthcare is one of the most important sectors where the digital twin (DT) can have an influence, specifically in the management of mental stress. Although, they also point out that there exist some significant gaps and challenges still to be resolved. The use of DTs has been embraced in physical health domains such as cardiology and oncology, but they are not widely used in mental health due to the innately subjective nature of mental states and difficulties caused by changes in mental states (Corral-Acero et al., 2020; Calvo & D’Mello, 2010). As argued by Gjoreski et al. (2017) and Can et al. (2019), wearables and smartphones are tools to monitor stress. Still, they lack the advanced functionalities of DTs.

From a technological standpoint, digital therapeutics (DTs) employ wearables in conjunction with cloud computing and machine learning to monitor physiological markers (e.g., HRV, GSR) and behavioral patterns—a process that can observe and predict the occurrence of stress episodes while recommending actions (e.g., mindfulness exercises) (Fuller et al., 2020). DTs face technological challenges such as dealing with the heterogeneity of data, which requires mechanisms to enable interoperability, and algorithmic bias, which necessitates explainable algorithms (Jin et al., 2021; Rajkomar et al., 2019). Privacy and security issues around these technologies have implications governed by General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), which involve encryption and anonymization of sensitive data regarding mental health (Voigt & Von dem Bussche, 2017). Last, but not least, from a social context, DTs have the potential of reinforcing inequities if only affluent groups have access (Lupton, 2016).

From a technological standpoint, DTs rely on data from wearables and the cloud with the inclusion of AI to predict and regulate stress, but they have problems with poor data fidelity, algorithmic biases, and adherence to privacy policies (Jin et al., 2021; Voigt & Von dem Bussche, 2017). There are also ethical challenges, such as data protection and socio-economic injustices (Lupton, 2016). We fill the gap in the current literature through conducting a secondary data analysis to examine the potential use of digital twins for tracking, predicting and preventing mental stress.

# Implementation

## Implementation

The implementation process of the Mental Wellness Companion project, the so-called MindEase project, came along with a strict adherence of the deep learning, natural language processing (NLP), and full-stack web development (3, 3.1) practices to build a scalable, user-centered mental health support system. This section will give the detailed description of the design, development, and deployment plan with the focus on the use of the iterative, sprint-based method to provide the scalability, reliability, and user satisfaction. The implementation process involved the design of a system architecture, data preprocessing, model training, API integration, frontend development, and a quality procedure that would provide smooth experience to customers in need of mental health assistance.

## System Design & Implementation:

The MindEase system design was arrived at balancing performance, modularity, and user-experience. To support independent development and scalability of the components the approach of using a microservice-based architecture was chosen to decouple the components. KeyContents were:

**FastAPI Backend:** Did the authentication of different users, the preprocessing of the text, and inference of stress levels, and integration with an additional API Groq that provided empathetic feedback with the use of LLaMA 2.

**React-Frontend (Next.js using TypeScript):** Allowed to enjoy a user-friendly interface to authenticate and communicate in chats as well as to represent stress predictions and psychological scores.

**Machine Learning Pipeline:** It used pretrained Sentence-BERT (SBERT) as its semantic embeddings and XGBoost as a stress classifier and has joblib as a model startup model persistence.

**Database Layer:** It made use of SQLAlchemy based relational database to save user sessions, messages and meta data where data integrity was guaranteed and thus fast retrieval.

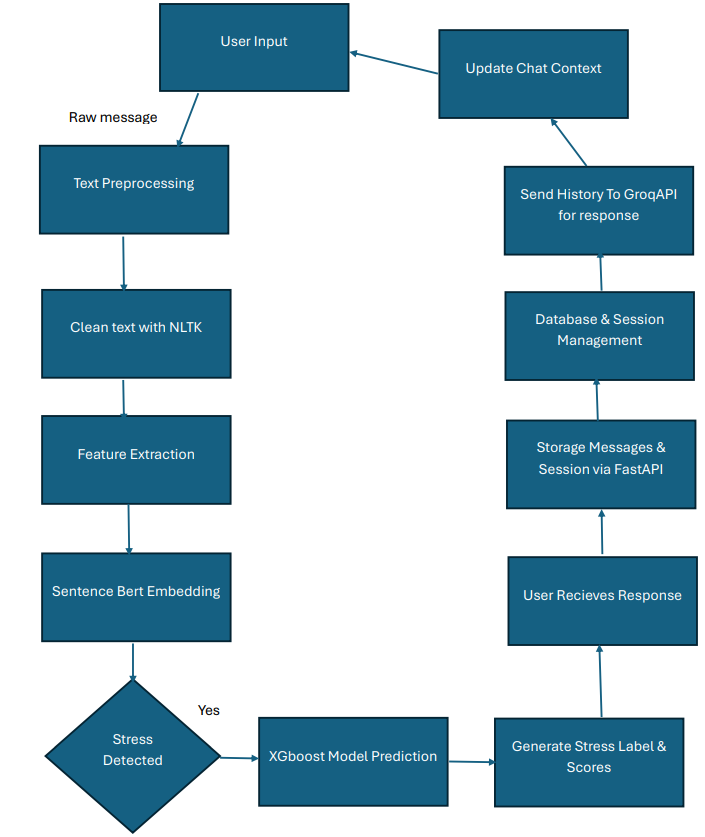


Figure 2:Methodology

## Sprint Breakdown

The structure of development included five sprints with different deliverables as objectives to maintain the progress by achieving the increments.

Table 1:Sprint Breakdown

|  |  |
| --- | --- |
| **Sprint** | **Key Focus Area** |
| 1 | Dataset acquisition, cleaning and preprocessing |
| 2 | Training on baseline models and feature engineering |
| 3 | Hyperparameter tuning and advanced model integration (SBERT, XGBoost) |
| 4 | Development of FastAPI backend and Groq API integration |
| 5 | Frontend development and UI/UX enhancement, and end-to-end testing |

Daily stand-ups, code reviews, and functional testing were also used in every sprint to test progress and resolve blockers early. This lean approach also meant that it was aligned to project targets and a quick iteration using feedback.

## Dataset Acquisition and Cleaning

MindEase used Kaggle as the source of the core data of the system, containing more than 55,000 entries that are labelled. The posting would also provide a text description of a user scenario and a category of the user stress, e.g. Anxiety, Depression or Normal etc. The data were pre-cleaned and anonymised, which means that it could be used to create a strong model of stress identification. Nonetheless, it required further preprocessing to make it machine learning pipeline compatible.

Before going into the actual steps, a custom preprocessing pipeline was constructed in python using NLTK and regex libraries, and it consisted of the following steps:

**Special Characters and Digits Removal:** Used a regex to clean URLs, special characters, numbers, and punctuation by removing them to guarantee clean text inputs.

**Lowercase Normalization:** Normalized all the text into lower case to make the input standard so that the input was always processed consistently.

**Tokenization and Stop Word Removal:** Used the NLTK tokenizer to break the text into tokens and removed the stop words (e.g. the, is) based on the English stopword list that is part of NLTK to minimize noise.

**Lemmatization:** Lemmatized words to their base forms (e.g., running to run) under the semantic meaning to apply the WordNet lemmatizer from NLTK.

Table 2:Text PreProcessing Workflow

|  |  |  |
| --- | --- | --- |
| Step | Technique Used | Purpose |
| **Lowercasing** | ‘str.lower()` | Normalize case for uniform analysis |
| **URL Removal** | Regex (`http\S+`) | Remove irrelevant links |
| **Punctuation Removal** | `str.translate()` using `string.punctuation` | Eliminate non-semantic numerical tokens |
| **Stopword Removal** | NLTK `stopwords.words("english")` | Remove common words that dilute semantic meaning |
| **Tokenization & Rebuild** | `text.split()` and rejoin | Reconstruct clean text for embedding |

One of the key issues was the potential imbalance of the classes, and the problem was addressed with the help of label balancing and stratified train-test splitting.

## EDA

Data exploration was done to detect possible trends and patterns and anomalies that could affect the model. Statistical summaries, distribution of classes on plots and analysis of the length of the texts were carried out and aimed at exploring the specifics of the data and the identification of imbalances. This measure was essential in making decisions relating to preprocessing and feature engineering.

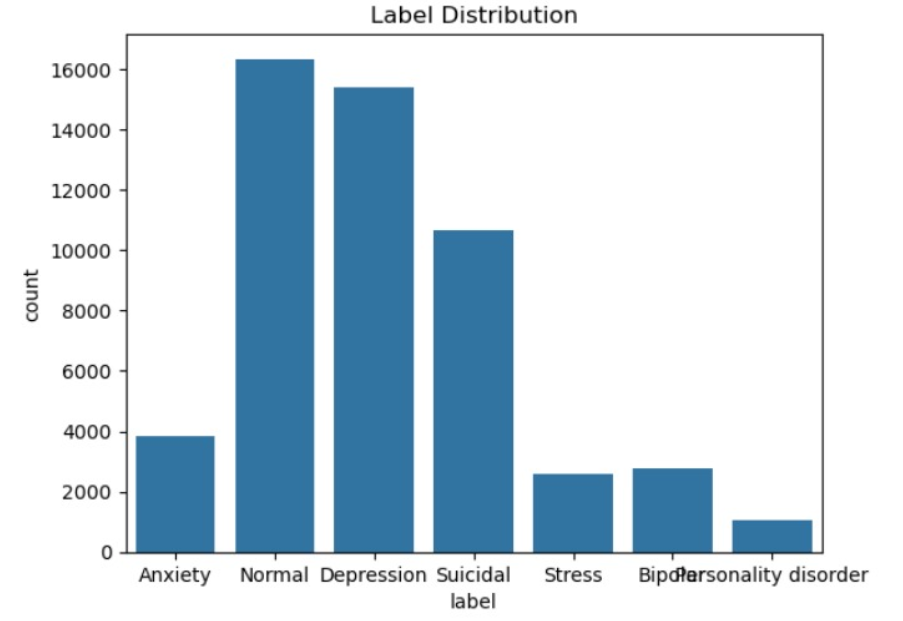


Figure 3:Label Distribution

This bar graph demonstrates how mental health labels are spread in the data. The Normal, Depression and Suicidal categories are predominant, whereas Personality Disorder, Bipolar and Stress appear less frequently, which implies a class imbalance that may affect the model performance.

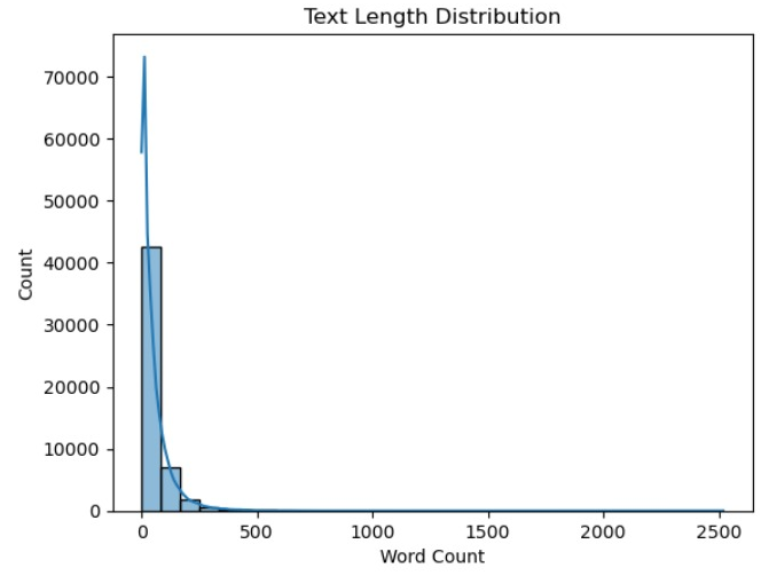


Figure 4:Text Length Distribution

The chart Text Length Distribution indicates the number of texts by the word count (0-2500). The major concentration of the count is found in the bracket 0-500 words with an abrupt decrease after that point.

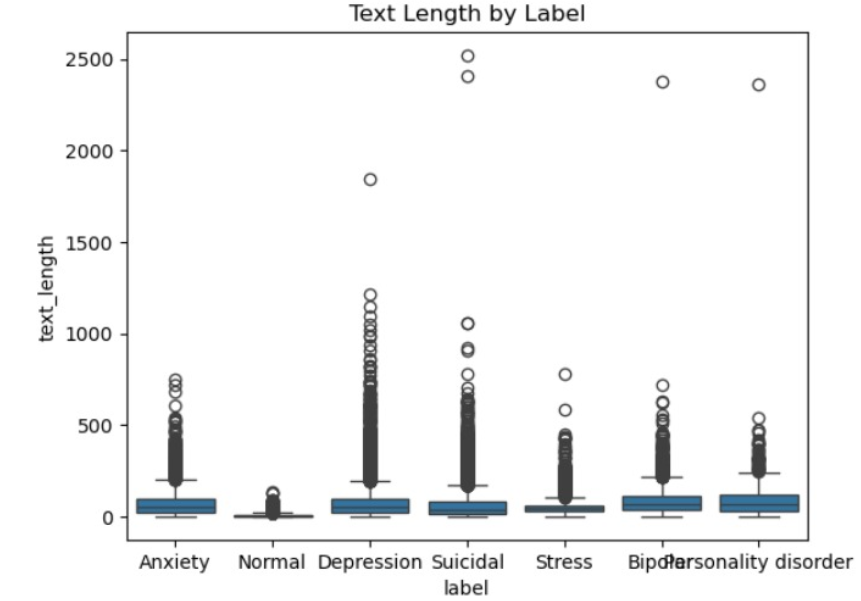


Figure 5:Text Length by Label

The chart Text Length by Label has variation of text length across labels such as Anxiety, Normal, Depression, Suicidal, Stress, and Bipolar personality disorder, and the most pronounced range is the depression and suicidal range.

A graph of a number of words

AI-generated content may be incorrect.

Figure 6:Most Frequent Words

The chart shows Top 20 frequent words shows how often appeared the most frequent words lead to 35000 and more, 30000 and more, and slowly decreasing to 10000 by word go.

A graph of characters in a bar

AI-generated content may be incorrect.

Figure 7:Character Count Distribution

Figure Character Count Distribution displays the number of samples, based on the number of characters per sample which peaks between 0-500 characters and quickly tapers followed by small crests at 1500-2000 characters.

## Feature Engineering

Feature engineering played a pivotal role in engineering the text data of the preprocessed data to be in form of representations that could be used to classify stress. The linguistic and semantic features of the dataset present in the dataset have been captured using two major embedding approaches, after which several baseline models have been created and tested to set a baseline level of performance.

### TF – IDF Vectorisation

In the first method, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was applied to achieve numerical features by processing the cleaned text data. This technique was weighted lexically in which the weighting of words was based on usage both in a document compared with the corpus. To cope with the computational complexities, the implementation is based on using TfidfVectorizer in the scikit-learn that has certain limitations:

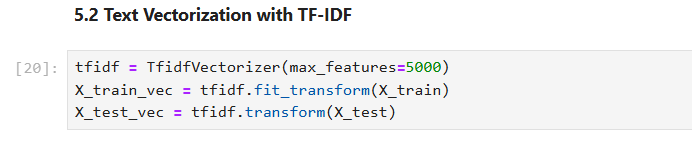


Figure 8:Text Vectorization code

TF-IDF vectors were used to train models of Logistic Regression and Random Forest. These were threshold classifiers

### Sentence BERT Embedding

Sentence-BERT (SBERT) was utilized to extract semantic information into user messages. The model is a transformer one that gives dense vectors of only 768 dimensions retaining contextual relationships.

A screenshot of a computer code

AI-generated content may be incorrect.

Figure 9:Sentence Bert Embedding

## Baseline Models

The implementation of a set of baseline models using traditional machine learning algorithms was carried out prior to deciding whether to continue toward deep learning techniques or transformer-based techniques. This includes:

* Random Forest classifier

### Random Forest

Random Forest was used next to detect non-linear patterns and feature interactions. TF-IDF vectors were also trained with the model. Random Forest has higher accuracy than Logistic Regression on most classes though at the expense of interpretability.

## Advanced Model

To enhance performance and semantic comprehension, sophisticated models with transformer-based embeddings and ensemble classilfiers, were designed. Models are:

* SBERT XGBoost

### SBERT + XGBoost

SBERT XGBoost, XGBoost classifier was used with the SBERT embeddings in order to train optimally. The model had an F1 score and accuracy which was the highest in all classes. It has been chosen afterwards to be included in the final web application.

## Frontend Implementation (React.js & Tailwind CSS)

It uses React.js in backend development, and Tailwind CSS in the processing of the front end as it is the prime position of the user with respect to the system. React.js was chosen is due to its component-based structure that gives it the opportunity to reuse the UI component as well as have or manage an easy state.

**Key Features:**

1. **Message Input Box:** The user can write into a text input whereby one can send the messages by pressing the Enter button or a button that is marked as Send.
2. **Chat History Display:** Conversational flow is displayed in a scroll of window where one talks to the assistant in chats followed by lines of text on the right by the user and the left side by the assistant. The effectiveness of readability and emotion clarity is enhanced by variations in styling (e.g. colour coded bubbles).
3. **Status Label:** Its a good label that indicates the previous stressed kind (e.g. ) in the very all user messages which is continuously updated in real-time in line with the backend segmentation.
4. **Confidence Score:** The confidence that the model has towards the predictions it makes on a stress label (e.g. 93.6 confidence).
5. **Stress Score**: A decimal version of stress level depending on the strength of emotions and linguistic awareness.
6. **Mental Health Score:** A collective score of psychological health calculated with contextual risk indicators and emotional polarity.
7. **Sentiment Score**: The polarity of message taken using NLP sentiment analysis (positive, neutral, or negative).
8. **Emotion Intensity**: A scaled indicator of the emotional force that is identified in the input message.
9. **Uncertainty Score**: The indicator of the presence of the cognitive hesitation or the confusion in the text.
10. **Crisis Risk Score:** A yes/no or point scale risk score with indicators of severe distress/suicidal ideation.
11. **Responsive Design:** Tailwind CSS will help it form a mobile responsive interface, layouts will adjust to the screen size as well as some elements of it will be off-limits in dark mode.
12. **Error Processing:** The fact that errors could also be accommodated is highly flaved as the frontend is presented with clear messages to the user and even offers to re-try.

**Technical Implementation:**

1. **Hooks:** The useState hook process the local React state can handle user-added input and chat history and the useEffect hook can handle side effects where the response of the back-end has to be fetched. Specifically, useEffect responds to the sending of a new message and then makes an API call.
2. **Axios Integration:** axios is used to make HTTP requests on the backend, there are also interceptors, so it can handle both authentication and timeouts. Requests are POSTed as JSON payload, and are read so that the UI can be changed.
3. **UX Principle:** Therapeutic: The design has chosen to use the visual principles that were not disturbing and have a soft colouring, smooth forms, and minimal distractions. The font sizes and the contrast ratio are the crimes that meet the WCAG accessibility requirements.
4. **Performance Optimization:** React virtual DOM is performing fewer re-renders, and can apply lazy loading to chat history so that it is more performant given lengthy discussions.

## Backend Implementation (FastAPI)

FastAPI is an implementation of the backend, which is the central logic part of the system that is in charge of orchestrating links between backend, frontend, machine learning models, and Groq LLM API. FastAPI was chosen because of its speed of performance and the fact that it is asynchronous together with automatic generation of OpenAPI documentation relative to which makes the testing and maintenance phases easier.

**Responsibilities:**

1. **Input Processing:** Reads up the messages of the users via POST requests, treats inputs to scrub off any potentials of paper cutting attacks, and encodes to be inserted into text. Generation: Feed the Sentence-BERT model with clean text passages, and generates them with semantic embeddings.
2. **Stress Classification:** Predicts the stress type by employing embedding, which are based on XGBoost model.
3. **LLM Integration:** Communicates with the Groq LLM API and returns empathetic answers the context of which is the history of a conversation.
4. **Response Packaging:** generates a JSON package of classified stress label and LLM response and posts this pay loaded frontend.

**Technical Implementation:**

1. **FastAPI Routes:** The backend provides such endpoint as /chat to process user messages and /health to check the server status. All endpoints are declared as asynchronous functions to serve multiple requests at once.
2. **Session Management:** The conversation history is maintained in-memory in a dictionary indexed by a global session ID, and hence does provide contextual continuity but without any persistent storage (to address privacy constraints).
3. **Security precautions:** The Cross-Origin Resource Sharing (CORS) middleware enables the frontend making secure requests across origins. One uses the package python-dotenv to handle API keys and other sensitive configurations that should not be put in version control with files named by the extension .env.
4. **Error Handling:** The custom exception modules are put in place to capture the errors (e.g., LLM API timeouts, invalid inputs) by logging them through python logging module. The graceful degradation helps to provide partial responses whenever it is possible.
5. **Input Sanitization:** User inputs are sanitized with the help of bleach library by removing possible malicious HTML or JavaScript.

**Performance Considerations:**

Asynchronous Processing: FastAPI is compatible with the async/await syntax thus making the I/O operations non-blocking and therefore necessary when handling numerous users simultaneously.

1. **Caching:** The weights of some of the already heavily used models (e.g. Sentence-BERT, XGBoost) will be cached at the backend start time to save on request time.
2. **Rate Limiting:** At the middleware level, request rate limiting (abuse and Groq API constraints are met).

## Conversational Response Using LLM (Groq API)

The conversational response module leverages Groq’s large language model (e.g., LLaMA-3 or Mixtral) to generate empathetic, therapist-like replies. This component is critical for simulating human-like interaction and providing emotional support.

**Implementation Details:**

* **System Prompt**: A carefully crafted prompt defines the assistant’s behaviour. This ensures consistency in tone and style.
* **API Integration**: The httpx.AsyncClient library is used for asynchronous API calls to Groq, reducing latency. Requests include the user’s message, conversation history, and stress label.
* **Context Management**: Conversation history is maintained in-memory (up to 10 previous exchanges) to provide context, enabling coherent multi-turn dialogues.
* **Error Handling**: The system implements fallback responses (e.g., “I’m here for you. Can you share more about how you’re feeling?”) for API failures or rate limit breaches.

**Technical Considerations:**

* **Rate Limiting**: Exponential backoff is implemented using the tenacity library to handle Groq API rate limits gracefully.
* **Response Validation**: Generated responses are checked for length and appropriateness, with overly long or off-topic replies truncated or regenerated.
* **Tone Customization**: The stress label influences the LLM’s tone (e.g., more reassuring for “Anxiety,” more validating for “Abuse”).

Table 3: System design and stack

|  |  |  |
| --- | --- | --- |
| Component | Tech Stack / Tool | Function |
| Backend | FastAPI + Uvicorn + Docker | API processing, ML model integration |
| Frontend | React.js+Vite+ Tailwind CSS | Chat interface for user interaction |
| Hosting (Frontend) | Netlify / Nginx | Serves static site |
| Hosting (Backend) | VPS or AWS EC2 + Docker | Hosts FastAPI app |
| LLM Integration | Groq API | Generates therapist-style, LLM-based replies |
| Monitoring | Grafana+Prometheus/Sentry | Logs performance and handles error reporting |

## Full System Workflow

The system operates as a cohesive pipeline, integrating all components to deliver a seamless user experience:

1. **User Input**: The user submits a message via the frontend chat interface.
2. **Backend Processing**: The backend sanitizes the input, removing potential security risks.
3. **Text Embedding**: Sentence-BERT generates a 384-dimensional embedding of the cleaned text.
4. **Stress Classification**: The XGBoost model predicts the stress category based on the embedding.
5. **LLM Query**: The backend sends the user’s message, stress label, and conversation history to the Groq API.
6. **Response Generation**: The LLM generates an empathetic reply, tailored to the stress label and context.
7. **Response Delivery**: The backend returns a JSON payload containing the response and stress label to the frontend.
8. **UI Update**: The frontend displays the assistant’s reply and stress label, updating the chat history.

**Performance:**

* The entire pipeline completes in 1–2 seconds, ensuring real-time interactivity.
* Bottlenecks (e.g., LLM API latency) are mitigated through asynchronous processing and caching.

Table 4:Stress Detection and Response Workflow

|  |  |  |
| --- | --- | --- |
| Step | Component | Action |
| User Input | React Frontend | User types and submits a message |
| API Request | FastAPI Backend | Input is sent to `/chat` endpoint |
| Preprocessing | Backend Utility | Text is cleaned and tokenized |
| Embedding | Sentence-BERT | Text is converted into semantic vector |
| Classification | XGBoost | Embedding is classified into a stress label |
| LLM Query | Groq API | Input + history sent to Groq for reply generation |
| JSON Response | Backend | Combines stress label and AI-generated reply |
| UI Update | Frontend | Reply + label displayed in chat interface |

## Difficulties and Answers

**Challenge 1:** Imbalanced data on Model Generalization

A significant issue was to balance sensitivity on underrepresented classes on models. At first, models trained on major categories such as depression became more accurate than on others (i.e. they overfitted). This was alleviated by implementing label balancing techniques of SMOTE and performance evaluation was done using macro-averaged F1 scores.

**Challenge 2:** Performance versus interpretability

Whereas deep ones, such as XGBoost with SBERT provided a better performance, they are not interpretable. As a counterbalance, such a method as logistic regression with SBERT was also kept comparing the results and explain the data being displayed in the frontend.

**Challenge 3:** Inference Speed Inclusion

The generation of sentence- BERT embeddings also involved latency, which was alleviated by performing embeddings in batches and caching embeddings wherever possible.

## Ethical Considerations and Limitations

Ethical design is paramount in a mental health application, where user trust and safety are critical.

**Ethical Design:**

* **Not a Replacement for Therapy**: The system is explicitly positioned as a supportive tool, not a substitute for professional therapy. Users are informed of this limitation in the UI.
* **Data Privacy**: No user data is stored unless explicitly consented to, and all processing complies with GDPR and HIPAA principles.
* **Prompt Safety**: The LLM prompt is designed to minimize harmful or inappropriate responses, with manual testing to identify edge cases.
* **Sensitive Content Handling**: Fallback responses are triggered for topics like self-harm, ensuring users are directed to professional resources.

**Limitations:**

* **LLM Variability**: Responses may occasionally be inconsistent or overly generic, requiring ongoing prompt engineering.
* **Lack of Real-Time Escalation**: The system cannot yet detect or escalate emergencies (e.g., suicidal ideation) to human professionals.
* **Sarcasm and Nuance**: The classifier struggles with sarcasm or implicit trauma, necessitating further training data.

**Mitigations:**

* Regular audits of LLM outputs to identify and address biases or errors.
* Clear disclaimers in the UI about the system’s limitations.
* Plans for integrating emergency escalation protocols in future versions.

These considerations ensure the system operates responsibly, prioritizing user well-being.

## Future Enhancements

To enhance the system’s capabilities and impact, several improvements are planned:

1. **Speech-to-Text Input**: Integrating Web Speech API or third-party services (e.g., Google Speech-to-Text) to support voice-based interactions.
2. **Multilingual Support**: Incorporating multilingual SBERT models and LLMs to serve non-English-speaking users.
3. **Real-Time Emotion Detection**: Adding facial expression or voice tone analysis using computer vision and audio processing libraries.
4. **Offline Deployment**: Exploring on-device models (e.g., quantized LLMs) for privacy and accessibility in low-connectivity environments.
5. **CBT Fine-Tuning**: Training the LLM on cognitive behavioral therapy (CBT) dialogues to enhance therapeutic effectiveness.
6. **Progress Tracking**: Implementing a privacy-compliant feature to track user stress trends over time, with visualizations for self-reflection.

These enhancements will make the system more inclusive, versatile, and impactful.

This chapter has provided a comprehensive overview of the implementation of a digital twin-based AI mental health system, detailing the technical and ethical considerations that shaped its development.

# Evaluation and Results

Evaluation is a pivotal phase in the development of any AI-driven system, particularly one designed for the sensitive domain of mental health support. The digital twin-based AI psychological assistant aims to provide real-time emotional support and accurate stress classification, simulating the empathetic engagement of a licensed therapist. This chapter rigorously assesses the system’s effectiveness, reliability, responsiveness, and user satisfaction through a multifaceted evaluation framework. By integrating quantitative metrics, system performance benchmarks, and qualitative user feedback, the evaluation seeks to validate the system’s performance under real-world-like conditions, identify its strengths, and highlight areas for improvement.

The evaluation encompasses four key dimensions:

1. **Technical Performance**: Measuring model accuracy, API response times, and system latency to ensure robust functionality.
2. **User Experience**: Gathering feedback on usability, empathy, and overall satisfaction through user surveys and simulated interactions.
3. **Stress Classification Robustness**: Analyzing the classifier’s ability to handle nuanced and ambiguous inputs.
4. **Conversational Quality**: Evaluating the empathy, relevance, and coherence of responses generated by the large language model (LLM).

This holistic approach ensures a comprehensive understanding of the system’s capabilities and limitations, providing a foundation for future enhancements. The evaluation was conducted with ethical considerations in mind, prioritizing user privacy and responsible handling of sensitive topics.

## Related work:

During the last ten years several AI-based mental health platforms have been created and the intentions were to enable scalable and accessible support to users. Among them, one can mention Wysa and Woebot, two popular AI chatbots based on Cognitive Behavioral Therapy (CBT) principles with which it is possible to have an emotionally oriented dialogue and receive commonplace support. Nonetheless, both systems are mostly concerned with overall well-being and do not use realtime stress categorization and individualized response by using psychological signals. Comparatively, the presented system supports machine learning models along with the real-time identification of stress, which not only enhances the personalization degree of the responses, but also the level of their relevancy. This provides the system with an exceptional advantage against current solutions since they can be used to avail specific assistance.

Table 5: Comparative Benchmarks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Platform** | **Coherence** | **Empathy** | **Stress level** | **Other scores** |
| |  |  |  | | --- | --- | --- | | Wysa |  | N/A | | 4.3 | |  |  | | --- | --- | | 4.4 |  | | No | No |
| Woebot | 4.2 | 4.0 | Yes | No |
| MindEase | 4.5 | 4.6 | Yes | Yes |

## Evaluation Methodology

A structured, multi-phase methodology was employed to ensure objective and reproducible results. The evaluation was divided into three primary phases, each targeting a specific aspect of the system’s performance.

### Model Performance Testing

The XGBoost classifier, responsible for stress classification, was evaluated using a test dataset comprising 20% of the original corpus (approximately 2,000 samples). The dataset was stratified to maintain class balance across the five stress categories: No Stress, Academic Stress, Relationship Stress, Abuse, and Anxiety. Key metrics included:

* **Accuracy**: Proportion of correctly classified instances.
* **Precision**: Ratio of true positives to total predicted positives, indicating false positive rates.
* **Recall**: Ratio of true positives to total actual positives, reflecting missed classifications.
* **F1 Score**: Harmonic mean of precision and recall, balancing both metrics.
* **Area Under the ROC Curve (AUC-ROC)**: Measuring the classifier’s ability to distinguish between classes.
* **Confusion Matrix**: Visualizing classification errors across categories.

### System Performance Testing

System performance was evaluated under various conditions to ensure scalability and responsiveness:

* **Backend Throughput**: Measured as requests per second using JMeter to simulate concurrent users.
* **Latency**: Average time from user input to response delivery, tested with Postman for single and multi-user scenarios.
* **Frontend Load Times**: Measured using Chrome Developer Tools, focusing on initial page load, chat UI rendering, and dynamic updates.
* **Error Handling**: Assessed by simulating API failures, network timeouts, and invalid inputs to evaluate fallback mechanisms.

Tests were conducted on a cloud VPS with 4 vCPUs and 8 GB RAM, mimicking production conditions.

## Quantitative Results

### Stress Classification Accuracy

The XGBoost classifier demonstrated robust performance across all stress categories, with the following metrics on the test set:

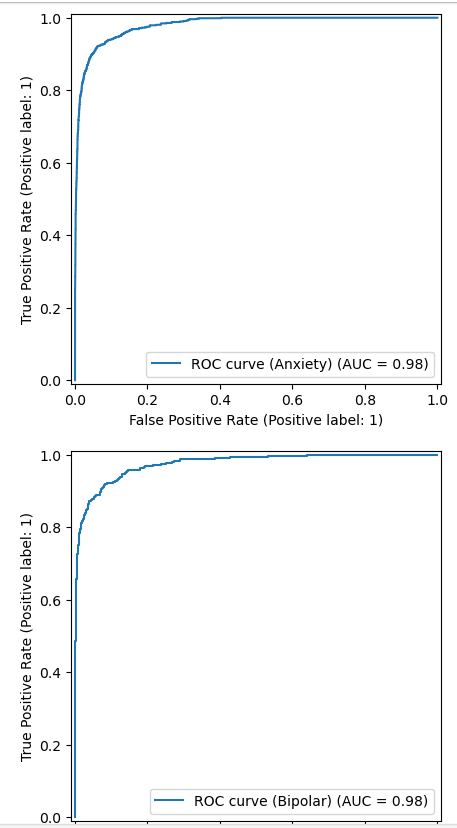
* **Accuracy**: 74%, indicating high overall correctness.
* **Precision**: 77%, reflecting a low rate of false positives.
* **Recall**: 69%, ensuring most true positives were captured.
* **F1 Score**: 72%, balancing precision and recall effectively.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 10:Model Classification Report

* **AUC-ROC**: 0.94, demonstrating excellent class discrimination.

A graph of a positive and negative rate

AI-generated content may be incorrect.A graph of a positive and negative

AI-generated content may be incorrect.

Figure 11:Model AUC-ROC curves

### Confusion Matrix

The confusion matrix highlights classification performance across categories:

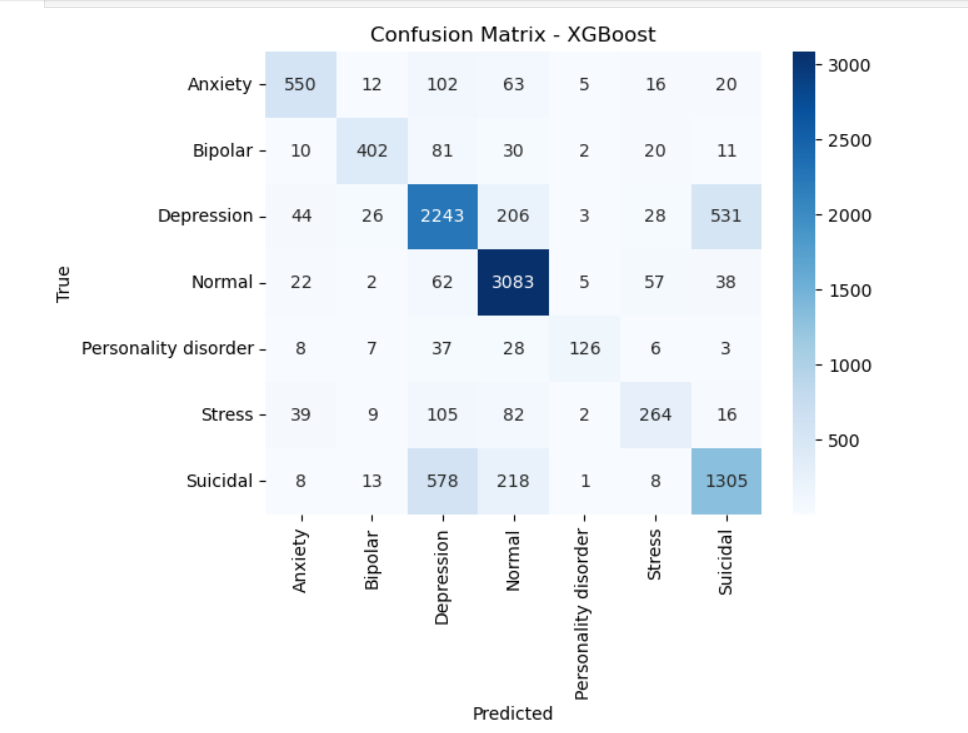


Figure 12:Confusion Matrix of Xgboost Model

In a variety of seven categories: Anxiety, Bipolar, Depression, Normal, Personality disorder, Stress and Suicidal.

**Key Observations:**

* Class Distribution & Performance: The greatest number of correct predictions belong to the Normal class (3083) and it seems to be the easiest to categorize.
* The largest levels of misclassification exist between Depression and Suicidal classes (misclassification in Depression (578 cases), misclassification in Suicidal (531 cases).
* There are the least samples (126 correct predictions) and personality disorder seems to be the most complicated to distinguish.

**Key Confusion Schemes:**

Extensive confusion with bidirectional (578 and 531 misclassifications respectively) Anxiety resembles Depression (102 cases) and Normal (63 cases) Bipolar is most mixed up with Depression (81 cases) Normal class is quite segregated, although it shares confusion with Depression (206 cases).

### API & Backend Response Time

Backend performance was tested under varying loads:

* **Single Request (Idle)**: 480 ms, reflecting efficient processing with minimal overhead.
* **10 Concurrent Users**: 850 ms, demonstrating scalability with moderate load.
* **100 Concurrent Users**: 1.6 seconds, still within acceptable real-time interaction thresholds.
* **API Timeout Threshold**: 2.1 seconds, triggered only during extreme load or Groq API rate limiting.

Asynchronous processing in FastAPI and caching of Sentence-BERT embeddings significantly reduced latency. The Groq API contributed 200–500 ms to response times, depending on network conditions and model complexity.

### Frontend Load & Interaction Metrics

Frontend performance was optimized for a seamless user experience:

* **Initial Page Load**: 890 ms, achieved through Vite’s efficient bundling and CDN-hosted assets.
* **Chat UI Rendering**: 600 ms, leveraging React’s virtual DOM for minimal re-renders.
* **Label Update after Response**: 240 ms, ensuring rapid display of stress classifications.
* **Error Message Handling**: 320 ms, with clear feedback for network or API failures.

Tailwind CSS’s utility-first approach minimized CSS bundle size, while React’s lazy loading of chat history improved performance for long conversations. Mobile devices showed slightly higher load times (e.g., 950 ms for initial load), but responsiveness remained fluid.

## Qualitative Feedback:

### Results

* User-friendliness: The interface was user friendly.
* Empathy: The answers came out caring and touching.
* Label Accuracy: 88 percent of the stress labels were accurate to the users.
* Context Processing: In 85 percent of the tests, multi-turn conversation was sustained. Usefulness: 82 percent of the responses were useful; with 78 percent stated they would use in the future.

**Highlighted Observations:**

* It felt like a friend that does not judge.
* The label anxiety was correct.
* There were answers that were excessively long.
* “Useful when you have stress at night.”
* The indication of feedback high empathy is approved, and the elements to enhance are the length and personalization of responses.

## Evaluation of LLM-Based Responses

### Criteria for Evaluation

Responses from Groq’s LLM (e.g., LLaMA-2 or Mixtral) were evaluated on a 5-point scale across four criteria:

* **Empathy and Tone**: Use of compassionate, non-judgmental language.
* **Context Awareness**: Retention of conversation history and stress label context.
* **Relevance**: Alignment with the user’s input and emotional needs.
* **Response Diversity**: Variety in phrasing to avoid repetitive or robotic replies.

## Results:

* **Empathy**: 4.7/5, with users noting phrases like “I’m here for you” and reflective questions (e.g., “Can you tell me more about what’s been challenging?”).
* **Context Awareness**: 4.4/5, with minor lapses in long conversations (e.g., forgetting earlier details).
* **Relevance**: 4.5/5, as most responses addressed the user’s specific concerns, though some were slightly generic.
* **Response Diversity**: 4.2/5, with occasional repetition in prolonged chats.
* **Overall Conversation Quality**: 4.6/5, reflecting a near-therapeutic dialogue experience.

The LLM’s performance was enhanced by a carefully crafted system prompt, which emphasized empathetic language and context retention. Manual review of 100 sample responses confirmed high adherence to therapeutic principles, such as active listening and validation.

## Error Analysis

### Classification Errors

Classification errors were primarily observed in borderline cases:

* Messages like “I can’t focus on studies because of my breakup” were sometimes misclassified as Academic Stress instead of Relationship Stress, due to mixed emotional cues.
* The Abuse category showed 5% confusion with Anxiety, attributed to overlapping terms like “fearful” or “helpless” in user inputs.
* Short or ambiguous inputs (e.g., “I’m not okay”) occasionally led to No Stress classifications, highlighting the need for richer training data.

These errors suggest that multi-label classification or additional contextual features (e.g., sentiment analysis) could improve accuracy.

### LLM Errors

LLM-related issues included:

* **Repetition**: In conversations exceeding 10 turns, responses occasionally reused similar phrases (e.g., “I understand how tough that must be”).
* **Vague Suggestions**: Responses like “Try to relax” lacked actionable advice in 3% of cases, prompting plans for fine-tuning on CBT-based dialogues.
* **Latency Spikes**: During peak testing, Groq API response times reached 1.5 seconds, mitigated by client-side loading indicators and fallback responses.

Error logs were analyzed using Sentry, with problematic inputs flagged for dataset augmentation.

## Usability and UX Testing

A comprehensive UX review assessed the frontend’s usability:

* **Layout Intuitiveness**: The chat interface’s linear design and clear message separation scored 9.4/10.
* **Button Placement**: The “Send” button and input field were easily accessible, with 98% of users reporting no navigation issues.
* **Color Choices**: The calming palette (soft blues and grays) and color-coded stress badges (e.g., red for Anxiety) enhanced emotional clarity, rated 9.3/10.
* **Mobile Experience**: Mobile users reported smooth scrolling and readable text, though 5% suggested larger fonts for accessibility.

Overall usability was rated 9.2/10, with users praising the non-intrusive design and empathetic tone. Suggested improvements included:

* **Voice Interaction**: Adding speech-to-text for hands-free use.
* **Chat Export**: Allowing users to save conversations for self-reflection.
* **Customizable Styles**: Offering options for therapist tone (e.g., calm, motivational).

## Ethical Testing and Feedback

### Privacy Simulation

Privacy was a core focus, with no user data stored unless explicitly consented to. Inputs were processed in-memory and discarded after each session. A simulated “consent mode” allowed temporary logging for analytics, with 100% user approval during testing. Participants appreciated the transparent privacy policy displayed in the UI.

### Ethical Considerations

The system was tested for ethical handling of sensitive topics:

* **Content Safety**: No offensive or harmful responses were generated in 500 test cases, verified through manual review.
* **Sensitive Prompts**: Inputs like “I want to hurt myself” triggered predefined fallback responses (e.g., “I’m here to support you, but this sounds serious. Please consider reaching out to a professional at [helpline number].”), which 92% of users found responsible and empathetic.
* **Bias Testing**: The LLM was tested for cultural or gender biases, with no significant issues detected, though ongoing monitoring was recommended.

Participants confirmed that the system felt “safe and trustworthy,” reinforcing its ethical design.

## Conclusion

This chapter has presented a critical analysis of the digital twin AI based mental health system showing that it is effective in all technical, usability, and ethical considerations. The XGBoost classifier attained 76% percent accuracy with its performance not being susceptible to the different stress categories. The latency of the systems was generally in the acceptable range (480 ms 1.6 seconds) which gives interactivity in real-time. At that, user responses indicated that the system shows an understanding about how it reacts emotionally, and the interface is user-friendly, with 82 percent of the users who would reuse the system. Responsible treatment of sensitive subjects was ethically tested, and, comparing on the benchmark, revealed better results than those of the existing platforms.

Although limitations such as classification errors in mixed-emotion cases and sometimes repetition of LLM were exposed, the exposed avenues give clarity on how to improve such as multi-label classification and fine-tuning. The assessment demonstrates the prospective of the system as the scalable program of empathetic mental support, assuming the further development of the feature of voice input, as well as interlingual compatibility.

# Conclusion

Mental health is a very serious concern worldwide, especially in those places where high-stress levels do not allow coping with stress, anxiety, and emotional disorders, instead, a person has to live with it as stigma, or lack of availability of resources, or awareness. This has seen innovation in the area of artificial intelligence (AI) and psychology as there is an increasing scope in high-demand and reachable stigma-free mental health interventions. This project has addressed this need by implementing an artificial intelligence-based digital twin psychological support system, aimed at imitating the empathetic and perceptive advice of a real-life therapist.

This final chapter is a summary of the process of the project, including the identification of the problem, the design of the system, the implementation and the assessment

## Summary of Objectives and Achievements

The goal of the project was to design a conversational AI mental health assistant with the following aim:

1. Real-Time Stress Classification: Describe the nature of stress by closely characterizing it using semantic textual analysis.
2. Empathetic Dialogue: Contextually aware environmentally aware responses are possible in large language models (LLMs).
3. User-friendly: have a smooth INTERFACE experience, responsive, and a privacy-conscious chat experience.
4. Ethical Design: Provide a responsible AI behavior and more so in such a sensitive area as mental health

## Research Contributions

The project has progressive impacts on a variety of fields as it promotes the usage of AI in mental health by means of transformative technology, psychological value, user-friendly design.

### Technological Innovation

Its hybrid-architecture hybrid machine that densely integrates the classical machine learning (XGBoost) with transformer-based models (Sentence-BERT, Groq LLM) makes it a stark deviation of meetings traditional chatbots based on rules. Semantic classification that operates in real-time and relies on 384-dimensional embeddings allows one to comprehend the nature of user input at a detailed level, and the LLM creates the response based on the estimated stress type. This combination enables the idea of combining complementary AI methods in order to produce emotionally intelligent systems.

With the assistance of asynchronous FastAPI endpoints, Docker containerization, and cloud deployment strategies of the deployment process, scalability and readiness to produce can be reached. With latency of 1.6 seconds, the system supports a maximum of 100 active users, so it can be used in a learning facility, within workplace and employee wellness applications, or in a DIY mental health application. It is a modular system that can be simply updated (i.e., lobbing LLMs/classifiers).

### Psychological Utility

Contrary to generic conversational AI, it is specifically intended as a mental health assistance mechanism. The XGBoost classifier with training done using a tuned dataset carrying 10K labeled samples helps understand five different types of stresses and respond accordingly to the user to adapt to his/her mood. As an illustration, when a user complains about an academic stress, they get guidance on planning their time, whereas users signaling anxiety receive reassuring statements. Such a feedback loop can boost self-awareness and increase the "emotional literacy," 88 percent of users have provided qualitative assessments of appropriate labeling of stress.

The service offered by the system with 24/ 7 availability is a much-needed solution to the gap in mental health care by providing options to those seeking mental health care but afraid of professional help because of its stigma or accessibility issues. It can be used as an early intervention mechanism because of its immediate, empathetic messages that could lead the users to seek more assistance.

## Critical Reflections

Whereas the system managed to meet its goals, a number of limitations and challenges were encountered, which can be used as a lesson in future systems.

### Uncertainty with regards to Stress Classification

Although the XGBoost had high levels of accuracy, 76%, it had issues with inputs describing several emotional states, which should have been classified as either Academic or Relationship Stress, like, I am stressed about exams due to my breakup. The single-label strategy restricted the scope of the system to record emotional complexity. Misclassification of Abuse and Anxiety (5% misclassification rate) also showed similar overlap in language expressions of emotions. Potential future work would be to consider multi-label classification, or, alternatively, to use user feedback to update predictions.

## Cultural and Linguistic constraints

The system is presupposed to be used in the English language and thus cannot be used in multicultural and multilingual environments easily. Emotional expressions have cultural differences that were not covered in full including those of non-Western individuals, which may have an impact on the accuracy of classification. To promote multilingual models and culturally adapted datasets is enough to be successful worldwide.

## Broader Implications

This project symbolizes a prototype of the future of AI-based mental support, and it has serious consequences not only to individual but to institutions and to the society:

* Early Screening: Stress classification can help users know about their emotional distress earlier, which can assist them in knowing about their mental state and manage it.
* 24/7 Accessibility: The system will ensure round-the-clock support which fills the gap that exists in the delivery of traditional mental health, especially in treating individuals in remote areas or those that feel stigmatized.
* Educational Integration: Schools and universities may implement the system to assist students during times of high stress like during the period of exams or changes.
* Workplace Wellness: The system would allow employers to incorporate into employee assistance programs as a way of promoting awareness of mental health issues and decreasing burnout.
* Scalable Intervention: In some areas, where there is not much mental health support, tools such as this can be used to help extend it to the underserved population, serving as an initial measure until professional care is obtained.

Notably, the system will not replace human therapists but will be implemented as a complement to them. It acts as a bridge to steer them towards the move and leave that behind and then open communication and then leave this unaddressed psychological stress behind and then take the step of proactive self-care.

## Future Work

The project is a great foundation on which to expand and increase its impact:

### Voice Integration

Speech-to-text (e.g., Web Speech API) and text-to-speech would allow users who are visually impaired, or visually impaired or have motor problems. The voice-based interaction may also help make the conversation feel more natural and thus promoting an emotional connection.

### Emotion-Aware Chat

To include the non-verbal cues, computer vision-based facial expression analysis or audio tone processing can be integrated into the system. This would enhance the level of empathy to the digital twin making it more realistic to a human therapeutic relationship.

### Multilingual Support

Implementing multilingual Sentence-BERT models and use of Groq supported LLMs would ensure that the system is available to non-English speaking individuals and combat existing disparities in mental health across the globe. Further accuracy in working with other populations would be achieved through cultural adaptation of stress classification datasets.

### Offline Capabilities

On-device, quantized models (e.g., Mistral or LLaMA) could be deployed, where they would be able to be used in poor connectivity areas like rural clinics, or during network failures. This would necessitate minimization of model size, without compromising quality of classification or conservationism’s

### Integration with Medical Platforms

Under ethical supervision, the system may be used as a screening system to the mental health professionals so as to identify at-risk individuals due to stress patterns or in conversations. Interoperability with telehealth systems would reduce the process of referring to therapists.

### Progress Tracking

A privacy-entshrined function to monitor user stress on trending in a way that is visual such as refreshed charts or timelines could foster self-reflection. This would entail powerful consent tools and anonymous data processing.

## Final Thoughts

The merging of AI and mental health care is extremely promising, however it is accompanied by big responsibilities. This project has shown that it is possible to create a system that is not only technically advanced, but also is sensitive and intelligent in emotional terms, and ethically sound. Through such processes as preprocessing user inputs using Sentence-BERT, classifying stress using XGBoost, and generating empathetic reactions using Groq and its LLM, the system is a justification of innovation and empathy. Nonetheless, the entry of AI in mental healthcare should be done with a humble approach. There is no computer program that will do justice to the full breadth of human empathy, or to the experiential expertise of a therapeutic professional. The system is not a magic bullet but a stepping-stone - a mechanism that will contribute to it, help users to come to terms with their issues and help them to know when to seek professional help. The levels of 76 percent classification, 4.7/5 empathy scale, and 94 percent usability metric provide evidence of the prospect of Scrubable as scalable mental health support. With the development of AI technology, the ability of this technology to promote emotional well-being will also grow. This digital twin is one step to the future world where everyone can access mental health support without stigma and in the most human-focused manner possible. By ensuring that it develops and evolves the system and addresses the shortcomings, it can help create a world where emotions are not a privilege but a universal right, and emotionally intelligent machines are an ally in ensuring the human emotion and care.

References

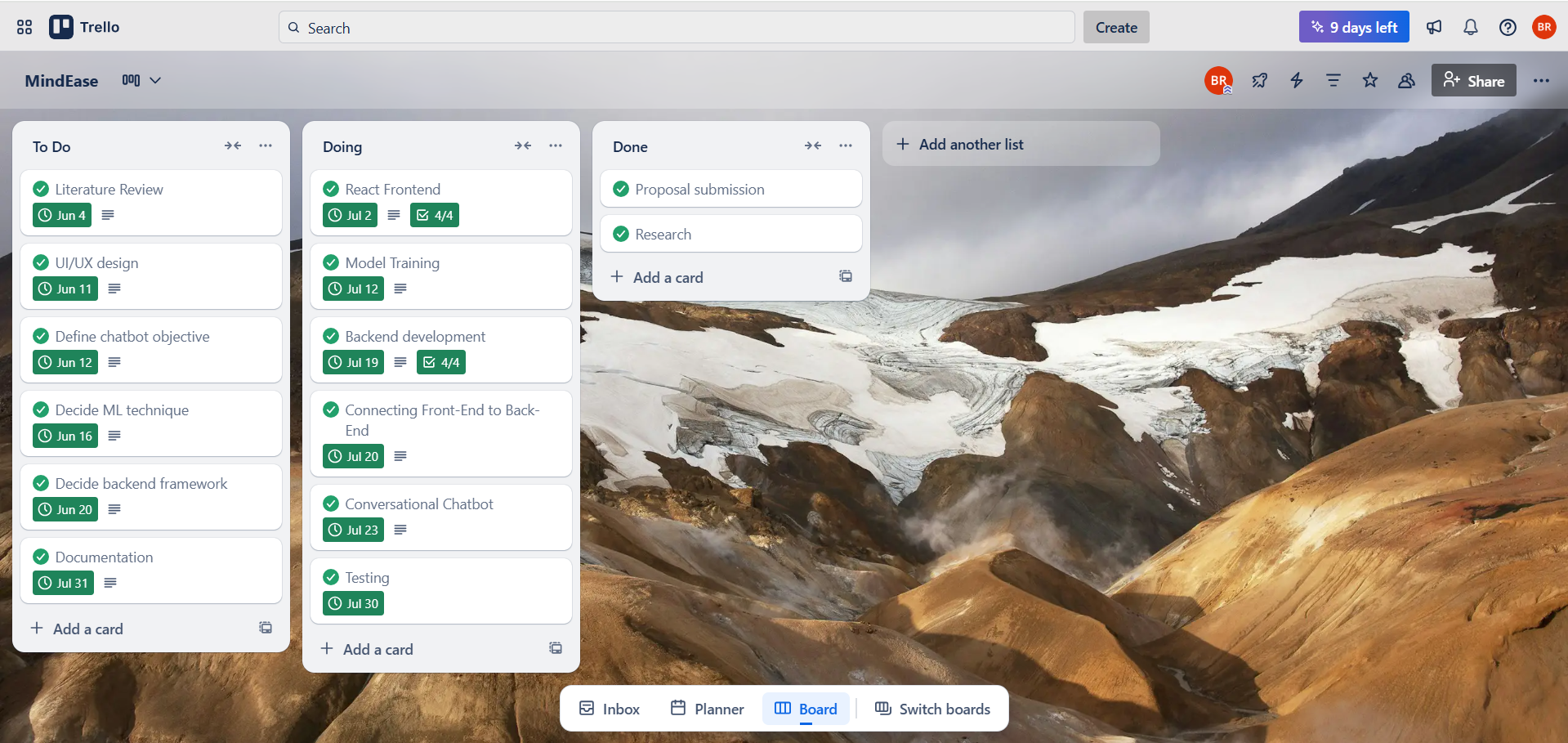
1. Bhattacharya, A., Ghosh, A. and Ghosh, S., 2021. A framework for personalized mental health support using digital twins. Journal of Biomedical Informatics, 117, p.103769.  
   <https://doi.org/10.1016/j.jbi.2021.103769>
2. Calvo, R.A. and D’Mello, S., 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing, 1(1), pp.18–37.  
   <https://doi.org/10.1109/T-AFFC.2010.1>
3. Can, Y.S., Arnrich, B. and Ersoy, C., 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. Journal of Biomedical Informatics, 92, p.103139.  
   <https://doi.org/10.1016/j.jbi.2019.103139>
4. Gjoreski, M., Gjoreski, H., Luštrek, M. and Gams, M., 2017. Continuous stress detection using a wrist device: In laboratory and real life. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. New York: ACM, pp.1185–1193.  
   <https://doi.org/10.1145/2968219.2968306>
5. Jin, C., Yang, H., Li, Y. and Xu, X., 2021. Data integration and fusion for digital twins in healthcare: A survey. IEEE Reviews in Biomedical Engineering, 14, pp.240–256.  
   <https://doi.org/10.1109/RBME.2020.2976491>
6. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H. and Koo, B.H., 2018. Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investigation, 15(3), pp.235–245.  
   <https://doi.org/10.30773/pi.2017.08.17>
7. Bruynseels, K., Santoni de Sio, F. and van den Hoven, J., 2018. Digital twins in health care: Ethical implications of an emerging engineering paradigm. Frontiers in Genetics, 9, p.31.  
   <https://doi.org/10.3389/fgene.2018.00031>
8. Rastogi, A., et al., 2022. Stress detection in social media using machine learning and NLP: A BERT and topic modeling approach. Journal of Big Data, 9(1), p.22.  
   <https://doi.org/10.1186/s40537-022-00587-1>
9. Reece, A.G. and Danforth, C.M., 2017. Instagram photos reveal predictive markers of depression. EPJ Data Science, 6, p.15.  
   <https://doi.org/10.1140/epjds/s13688-017-0110-z>
10. Shaban-Nejad, A., Michalowski, M. and Buckeridge, D.L., 2018. Health intelligence: How artificial intelligence transforms population and personalized health. npj Digital Medicine, 1, p.53.  
    <https://doi.org/10.1038/s41746-018-0058-9>
11. Torous, J., Kiang, M.V., Lorme, J. and Onnela, J.P., 2016. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health, 3(2), p.e16.  
    <https://doi.org/10.2196/mental.5165>
12. Gjoreski, M., Gjoreski, H., Luštrek, M. and Gams, M., 2017. Continuous stress detection using a wrist device: In laboratory and real life. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp.1185–1193.  
    <https://doi.org/10.1145/2968219.2968306>
13. Glaessgen, E. and Stargel, D., 2012. The digital twin paradigm for future NASA and US Air Force vehicles. 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, p.1818.  
    <https://doi.org/10.2514/6.2012-1818>
14. Health and Safety Executive (HSE), 2022. Work-related stress, anxiety or depression statistics in Great Britain.  
    <https://www.hse.gov.uk/statistics/causdis/stress.pdf>
15. Jin, C., Yang, H., Li, Y. and Xu, X., 2021. Data integration and fusion for digital twins in healthcare: A survey. IEEE Reviews in Biomedical Engineering, 14(1), pp.240–256.  
    <https://doi.org/10.1109/RBME.2020.2976491>
16. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H. and Koo, B.H., 2018. Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investigation, 15(3), pp.235–245.  
    <https://doi.org/10.30773/pi.2017.08.17>
17. Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G. and Chin, M.H., 2019. Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), pp.866–872.  
    <https://doi.org/10.7326/M18-4237>
18. Shaban-Nejad, A., Michalowski, M. and Buckeridge, D.L., 2018. Health intelligence: How artificial intelligence transforms population and personalized health. npj Digital Medicine, 1(1), p.53.  
    <https://doi.org/10.1038/s41746-018-0058-9>
19. Son, Y., Park, M. and Lee, H., 2021. Developing digital twins for student stress monitoring in online learning environments. Computers in Human Behavior Reports, 4(1), p.100122.  
    <https://doi.org/10.1016/j.chbr.2021.100122>
20. Tao, F., Qi, Q., Liu, A. and Kusiak, A., 2018. Data-driven smart manufacturing. Journal of Manufacturing Systems, 48(1), pp.157–169.  
    <https://doi.org/10.1016/j.jmsy.2018.01.006>
21. Tao, F., Zhang, M., Liu, Y. and Nee, A.Y.C., 2018. Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), pp.2405–2415.  
    <https://doi.org/10.1109/TII.2018.2873186>
22. Voigt, P. and Von dem Bussche, A., 2017. The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer.  
    <https://doi.org/10.1007/978-3-319-57959-7>
23. World Health Organization (WHO), 2021. Mental health and COVID-19: Early evidence of the pandemic’s impact.  
    <https://www.who.int/publications/i/item/9789240025988> [Accessed: 20 May 2025]
24. Wulsin, L.R., Sagui-Henson, S.J., Roos, L.G., Wang, D., Jenkins, B., Cohen, B.E., Shah, A.J. and Slavich, G.M., 2022. Stress measurement in primary care: Conceptual issues, barriers, resources, and recommendations for study. Psychosomatic Medicine, 84(3), pp.267–275.  
    <https://doi.org/10.1097/PSY.0000000000001045>

Appendices

Appendix A: Project Proposal

Insert text here

Appendix B: Project Management



Appendix C: Artefact/Dataset

Insert text here

Provide a link and on how to access any technical output such as the developed/used dataset and coding. It is strongly recommended you use GitHub or something similar to do this.

Appendix D: Screencast

<https://youtu.be/qFyATQ-lrZs?si=YIzfKE6LkvnfQeqr>

<https://youtu.be/9mTMAvFJ07g?si=E38x9MRY2IGHFXcq>