Integrating a hybrid model for a short term Mental Health prognosis

MSc Research Project
Msc Data analytics

Hariharan Vijayan
Student ID: x23404507

School of Computing
National College of Ireland

Supervisor: Furqan Rustam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Hariharan Vijayan
Student ID:	x23404507
Programme:	Msc Data analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Furqan Rustam
Submission Due Date:	11/12/2025
Project Title:	Integrating a hybrid model for a short term Mental Health prognosis
Word Count:	7608
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Hariharan Vijayan
Date:	11th December 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Integrating a Hybrid model for a short term mental health prognosis

Hariharan Vijayan
x23404507

Abstract

Increased rates of stress, anxiety, and depression have resulted in the need to develop mental health screening instruments that would be available to more individuals in a shorter period of time. Conventional evaluation techniques rely on doing it manually and using stagnant questionnaires which require time to respond to and are hard to customize to particular requirements and frequent changes in the mental state which only occur over days or weeks may be overlooked. These issues suggest the usefulness of automated mechanisms that are capable of processing mental health data, and also helping people gain more interesting and personalized means of considering their wellbeing. The proposed study presents a prototype that involves two steps, which include existing machine learning methods and a reflective questioning system fueled by a large language model. By using a large questionnaire data set comprising of DASS questions and demographics, the system will compute the score of depression, anxiety, and stress and categorize each participant into three risk groups. To predict such levels of risks, several supervised learning models such as Random Forest, Gradient Boosting, AdaBoost and XGBoost were trained and tested. The models performed well with the weighted F1 scores of 0.90-0.94, indicating that all the approaches have strong classification. During the second stage, the system takes the level of predicted risk and relevant data features together to produce customized reflective questions and brief narrative understanding with one of the large language models. Although this prototype is not the one that will be used in clinical diagnosis and treatment, it demonstrates the effectiveness of predictive modeling and AI which can be conversational as the means of making digital mental health tools engaging

1 Introduction

The challenges of mental health have grown enormously in the societies as a great number of people report stress and anxiety, as well as depression symptoms. It is significant to recognize emotional strain at its initial stages. With the effective timing of the recognition, one might prevent the issues or the challenges and assist individuals in seeking help before their health condition worsens. Conventional mental health assessment procedures are based on structured clinical interviews and standardised questionnaires, which give useful information at a single time. It is important to have real time and continuous monitoring techniques that can record dynamic emotional patternsBhaganagare et al. (2025). The development of digital technologies has provided opportunities to develop

supportive tools that could be used to make people learn about their emotional patterns in nonclinical environments. Empirical research has shown that data in structured questionnaires can reveal trends associated with emotional vulnerability and that new methods of natural language interaction offer both potential and major design challenges in mental health assessment Cai et al. (2025). Recent research also showed that alterations in behaviour patterns gathered by the personal devices over short periods, may be used to forecast changes in the symptoms of depression and anxiety, as well as in changes in the mood in the nearest future Daga (2025). Based on these observations, more recent randomized controlled trials have revealed that digital interventions that combine wearable physiological data with mobile applications can reduce depressive and anxiety symptoms, and hence, produce substantial clinical benefits through personalized and data driven therapyFatouros et al. (2025).Simultaneously, advances in massive conversational systems have made the computer programs capable of producing personalised interactions that facilitate the process of reflective thought and prompts users to investigate their emotional lives in a natural and non judgemental fashion Choi et al. (2024). Collectively, these trends are a basis of hybrid systems that combine organized quantitative evaluation and conversational qualitative interpretation.
The current research is based on these developments by coming up with a two stage hybrid prototype. At the first stage, the supervised computational models are trained to predict the individual current level of emotional risk on the basis of the available questionnaire data on their responses. The second stage involves a conversational Large language model on the assessment results, alongside a knowledge base, to produce personalised reflective questions and interpret answers made by the user. The report is of value in four ways. It trains a monitored computational model to approximate existing emotional danger with extensive information of questionnaire replies (DASS dataset),It integrates a streamlit interactive system generating personalised reflective questions in accordance with an emotional profile of a user. It analyzes the potential to predict the short term (7–14 days) mental health interpretation of responses posed by users, It considers the possibility of providing early emotional awareness through the use of digital mental health tools by integrating quantitative evaluation with conversational reflection.

1.1 Research Question

Can a hybrid machine learning and language model system utilize the data gathered via questionnaires and customized reflective conversation to approximate the present mental health risk of an individual and offer an interpretable short term prognosis of the following one or two weeks?

1.2 Objectives

Design a Model which estimates the Current Emotional Risk,Conversational Language Based Reflective System,Short Term Emotional Tendencies.Test a Hybrid Predictive and Reflective Support System.

The rest of this report has the following structure. The following chapter gives a literature review. The methodology chapter outlines the procedures of data preparation, the modelling method and the system architecture. The implementation chapter provides the construction of the prototype. The experimental results are given in the evaluation chapter. The final chapter is a summary of the key findings and recommendations on future research.

2 Related Work

Structured self report (PHQ 9) and clinical interview (Generalised Anxiety Disorder- 7 (GAD-7)) have been used as the most common evaluations of mental health problems relying on accessibility, cost effectiveness and strong psychometric validation. Depression Anxiety Stress Scales (DASS-21) is another highly used tool in the measurement of cluster of symptoms related to depression, anxiety, and stress. Regardless of these significant strengths, the commonality of all these instruments is their cross sectional nature, allowing only one of the temporal samples of the state of emotional distress, which does not allow us to represent the dynamics of the emotional state of a person and does not give us information about the risk pattern in the short term. The traditional questionnaires also have inherent limitations such as self report bias, recall errors and the social desirability bias.

2.1 Passive Sensing and Digital Phenotyping Technologies

Digital phenotyping has occurred as a revolutionary approach that derives measurements in inadvertent communication with smartphones and other digital devices as indicators of mental state Bufano et al. (2023). The feasibility of digital phenotyping toward predicting relapse or symptom exacerbation has been investigated in systematic reviews which have identified a broad set of behavioral and physiological characteristics such as mobility patterns, phone usage measures, heart rate, sleep quality, and physical activity levels which can generate digital phenotypes that can be used to support clinical evaluation and predict symptomatic relapse Bufano et al. (2023). Mass application has proven to be promising with a cross sectional study of 10,129 UK participants who used Fitbit sensors and self reported questionnaires providing R-Squared values of 0.41 and 0.31 respectively in predicting depression and anxiety Zhang et al. (2025).The changing environment does not confine to passive surveillance to include smartphone apps, virtual reality, and generative artificial intelligence systems Torous et al. (2025). The modern digital interventions in mental health manufacture an unparalleled flexibility and access, but the discipline needs more stringent methodological methods such as digital placebo control and improved transparency Torous et al. (2025).
Surveys on the artificial intelligence applications have critically examined the capabilities of diagnosis and monitoring and intervention, which show the possibility of enhancing diagnostic accuracy and providing personalized therapeutic interventions, although issues of algorithmic transparency and bias reduction still prevail Cruz-Gonzalez et al. (2025).The analysis of smart device ecosystems has identified key characteristics required in digital phenotyping, which include accelerator data, count of steps, heart rate data, and sleep data as the consistently useful ones in predicting mood disorders Jung et al. (2025). The sensing capabilities that are implemented as a smartphone show that behavioral patterns such as communication frequency, physical activity levels, and location variability are associated with the severity of the symptoms of depression and machine learning models can predict the occurrence of depression with moderate accuracy Terhorst et al. (2025). The analysis of smartphone sensor data within youth populations (12-25 years

old) found that GPS, accelerator, and screen time measures were potentially useful indicators Beames et al. (2024). The passive sensing approaches have proven to be useful in the case of university student population to predict levels of stress Shvetcov et al. (2024), with studies of multimodal digital phenotyping showing high sensitivity and specificity in detecting episodes of depression in patients with major depression Aledavood et al. (2025).

2.2 Predictive Modeling and Machine Learning

The high volume of digital mental health data sustained by rapidly growing digital data has facilitated machine learning processes in capturing non linear relationships in high dimensional questionnaire data. The most commonly used classifiers are the logistic regression, support vectors machines, random forests, gradient boosting algorithms, and 4 neural network structures. Empirical studies have proved that machine learning methods offer greater sensitivity and specificity and classifiers often have a higher accuracy (above 85-90 percent) in differentiating between low-risk and high risk subjects. Although these encouraging findings were achieved, the existing methods have the key disadvantages such as the lack of time modeling, the difficulty of result interpretation and the inability to combine the results in such a manner that they can be reflected on the user level.

2.3 Large Language Models and Conversational Systems

The use of digital mental health has become a proliferating field, which provides psychoeducation, mood monitoring, and emotional assistance on mobile apps. Modern applications are more and more based on large language models that can generate natural and context aware chats. Large language model architecture systems have shown promise in assisting individuals in expressing emotion and learning self regulatory coping mechanisms, and meta analytic data have shown moderate to high positive impacts on psychological distress.

2.4 Hybrid Predictive Reflective Systems and Research Gap

Despite the fact that the literature on digital mental health interventions with adaptive features does exist, none of the systems have fully applied machine learning risk estimation on structured questionnaire data combined with a large language model based reflective questioning. Traditional measures of assessment provide valid quantification and no predictive values. Machine learning models have the capability of converting the responses of questionnaires into probabilistic risk estimations but do not provide presentations of predictions in forms that are comprehensible by users. Conversational agents promote individual reflection, but do not have predictive grounding. The encoding of these dissimilar capabilities is one of the key areas in developing digital mental health technology to systems that merge measurement rigor, predictive ability, and engagement advantages of natural language interaction interfaces.

Table 1: Summary of Key Literature Highlighting the Existing Research Gap

Study	Data Type	Method	Contribution	Limitation
Lovibond & Lovibond (1995) Lovibond and Lovibond (1995)	Self-report questionnaire	DASS-42 scoring	Assessed depression, anxiety and stress scale,	provides no predictive information about future emotional states
Bhaganagare et al. (2025) Bhaganagare et al. (2025)	It reviewed multimodal AI systems	Systematic review	opportunities for AI to monitor mental health in realtime.	Does not propose hybrid predictive–reflective systems
Cai et al. (2025) Cai et al. (2025)	Natural language interaction data	Design analysis	Determines opportunities and challenges of NLI-based mental health assessment tools.	Focuses on challenges rather than solutions; no integration with ML risk prediction
Daga et al. (2025) Daga (2025)	Passive mobile sensing data	AI/ML on sensor data	Identifies mental health deterioration 1–2 weeks before with 77–86% accuracy.	Relies on continuous passive sensing; not based on structured questionnaires
Fatouros et al. (2025) Fatouros et al. (2025)	Wearable sensors + mobile app	Digital therapeutic intervention	Achieves 45% reduction in depression and 50% reduction in anxiety.	Treatment-focused; requires wearable devices
Zhang et al. (2025) Zhang et al. (2025)	Wearable device data (10,129 participants)	ML prediction models	Achieves $R^2 = 0.41$ for depression and 0.31 for anxiety through large-scale digital phenotyping.	Cross-sectional; lacks temporal modelling or reflective reasoning
Torous et al. (2025) Torous et al. (2025)	Multiple digital health modalities	Review of smartphone apps, AI, VR	Provides overview of digital mental health landscape and evidence gaps.	Identifies need for robust hybrid architectures but does not propose solutions
Cruz-Gonzalez et al. (2025) Cruz-Gonzalez et al. (2025)	AI applications review	Systematic review	Evaluates AI for diagnosis, monitoring, and intervention.	Highlights transparency and bias issues; lacks interpretable system designs
Terhorst et al. (2025) Terhorst et al. (2025)	Smartphone sensor data	ML for depression severity	Behavioural patterns correlate with depression severity enabling moderate prediction accuracy.	It focus on current prediction not future
Nemesure et al. (2021) Nemesure et al. (2021)	Electronic health records	ML prediction models	Demonstrates superior performance of tree-based ensembles for mental health prediction.	Depends on clinical EHR data; limited discussion of interpretability for end-users

3 Methodology

This chapter describes the design, development, and testing of the proposed hybrid system which approximates the present level of psychological risk of an individual and is able to generate an interpretable short term emotional prognosis. The methodology is divided into seven sections: description of the dataset, psychological distress scoring and target construction, feature engineering, machine learning modelling, language based reflective reasoning, system architecture and ethical considerations.

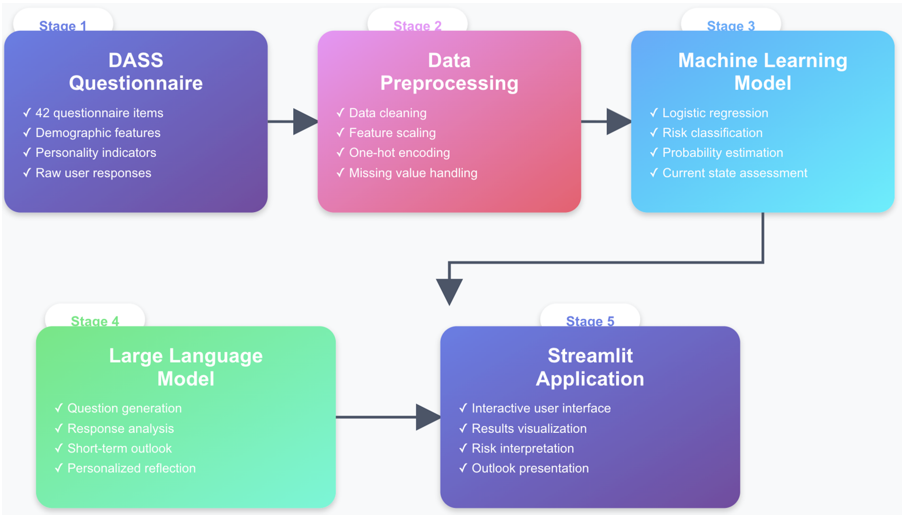

Figure 1: Overall System achitechture for Hybrid model

In Figure 1 , the architecture of the entire system of the hybrid predictive reflective model is provided. It describes the process of how raw questionnaire data is taken through preprocessing and machine-learning elements, and ultimately, combine with a large-language-model layer and Streamlit app to create interactive and personalised emotional insights.

3.1 Dataset Description

The system is constructed using a publicly available data set, which is constructed around the Depression Anxiety Stress Scales (DASS-42) which comprise about thirty-five thousand anonymised survey responses. The DASS-42 instrument is an instrument with a proved level of assessing depressive, anxiety, and stress symptoms in the clinical and non clinical groups, and the psychometric reliability of the instrument has been continually supported in the literatureLovibond and Lovibond (1995).The dataset contains answers to fourty two psychological items which are rated on the four point scale. Some variables that are optional like age, gender, level of education and nationality. Ten questions in a short personality scale which measured general personality traits.An item of sixteen vocabulary recognitions.Metadata used to capture timing data on the completion of the survey.The data is cross sectional.

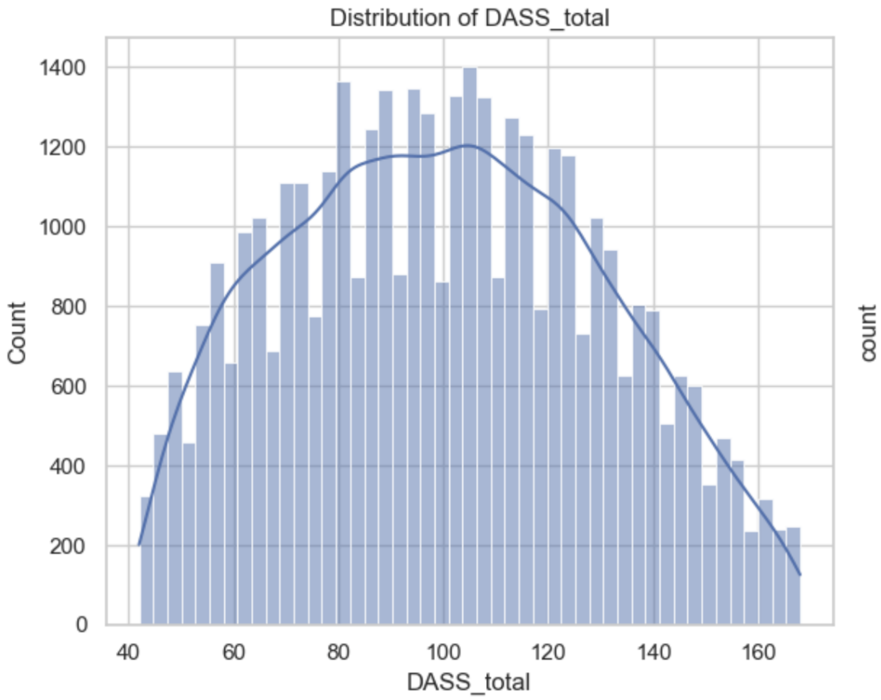

Figure 2: The Distribution of the total DAS scores across the dataset

All the individuals are observed just once and no time-varying measurements are observed. As a result, risk level estimation by the machine learning component of a system is only estimated at any given time. This information is then fed on by the language based reasoning component, which in conjunction with the written reflections of the person generates a short term future tendency without time series forecasting.In order to justify the future methodological decisions, a cursory exploratory analysis was done. There was a smooth continuous distribution of the DASS total scores and the resultant low , medium , and high risk classes were found to be well represented after quantile partitioning.

3.2 Construction of Psychological Distress Scoring and Target

Following the official rules of DASS-42 scoring, the raw item responses were converted into three psychological constructs namely: Depression, Anxiety, and Stress. A predefined subgroup of items was added up to come up with a subscale score in relation to each construct. These three subscale scores were then contrasted with the set cut offs and each individual participant was placed in one of the standard severity levels (Normal, Mild, Moderate, Severe, or Extremely Severe) of each construct.In order to facilitate the machine learning task, total distress index was calculated as the sum of the Depression, Anxiety, and Stress subscale scores of the individual participants. This overall DASS score was then converted to a three level risk labels named as low risk, High risk and medium risk.In practice, the overall DASS scores were broken into three categories based on sample quantiles such that the samples in each risk level were roughly one third of the total population. The three class risk label is the resultant variable that is used as the target variable of the classification models.

3.3 Feature Engineering

The initial data were condensed into two broad groups of predictors, namely, numerical and categorical features.This method correlates with the conventional feature engineering in mental health prediction, in which the demographic, behavioral, and psychometric features are systematically aggregated to improve model performance. Jacobucci and Grimm (2020)

3.3.1 Numerical Features

The indicators of timing of the surveys (taking time on introduction and main survey parts).Ten ratings of personality characteristics based on the brief personality questionnaire.There are 16 vocabulary recognition indicators. Quantitative demographic characteristics including age, the size of the household, and the size of the screen.When considered as continuous inputs, the responses of the forty two DASS are in the form of an item.

3.3.2 Categorical Features

The categorical characteristics are Demographic factors like gender, nationality, education level, and other demographic factors.Environment where one lives in (such as rural or urban),Political orientation,Marital status,Recruitment source and other descriptors of context.Data was of high quality in that records with much missing information were eliminated prior to analysis. With the rest of the data, the numerical features were normalised so as to have a similar distribution during training. Categorical data were one hot encoded into numerical data, so that each category observed in the data get a distinct binary indicator.In order to make sure that every type of feature obtained corresponding preprocessing, a pipeline of feature transformations was specified. This pipeline uses median imputation and standardisation of numerical variables, most frequent imputation and then one hot encoding of categorical variables. .

3.4 Machine Learning Modelling

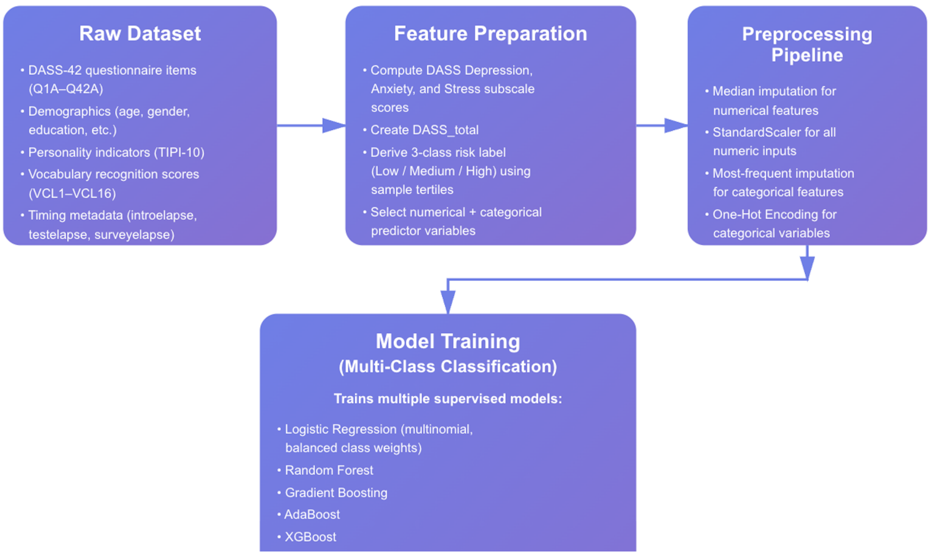

Figure 3: Overview of the Machine learning pipeline

Multi class classification experiment with a number of machine learning methods was performed under supervision. Multinomial logistic regression,Random forest classifier,Gradient boosting classifier,AdaBoost classifier,Support vector classifier,The nearest neighbour classifier,The extreme gradient boosting classifier.The choice of these models is due to prior studies of mental health risk prediction that have found that tree based ensemble and margin based classifiers frequently perform well on symptom and questionnaire dataNemesure et al. (2021). This project

aimed at the comparison of a variety of such models on the problem of risk labelling on three classes.In some of the runs, the extreme gradient boosting model was able to attain the greatest area under Receiver Operating Characteristic curve but logistic regression was selected as the main reference model. This option is driven by the fact that it is interpretable, its behaviour remains relatively constant with class imbalance and it is consistent with usual practices in psychological risk modelling.
Interpretability of logistic regression is more useful specifically in the clinical environment where the transparency of the decision making process is the key to trust and compliance with regulationTafti et al. (2021). An eighty twenty stratified sample split was used to divide the dataset into a test set and a training set. Stratification was done to maintain the low, medium, and high risk cases in the two sets in proportions.Stratified train test divisions play a vital role in mental health prediction tasks to maintain the distribution of uncommon cases but clinically significant cases and avoid biased model testing.Lundberg et al. (2018)Training pipeline is made of Preprocessing numerical and categorical input features ,placing the classifier on the transformed training data,generating class probabilities of three levels of risks on the held out test set.transforming these probabilities into the predicted labels through choosing the most likely class.The processing and division strategy applied to all models were the same so as to facilitate fair comparison.

3.4.1 Evaluation Metrics

There were two key metrics used to assess model performance:

- **Receiver Operating Characteristic Area Under the Curve (ROC-AUC):** The extent to which each of the models differentiates between the three risk categories was measured using a weighted one versus rest multi class ROC-AUC at various probability thresholds.One versus rest approach to multi class ROC-AUC will give a solid assessment framework that will consider the complexity of the task of differentiating between different levels of psychological threats. Luque et al. (2019)

- **F1-score:** The weighted F1-score was calculated to stabilize precision and recall of each class in case of class imbalance. The measure is especially appropriate in the mental health setting, where false negatives and false positives are both impactful

.

3.5 Reflective Reasoning over Language Model

When the machine learning model has determined the current emotional risk level, the second component generates a qualitative description of the changes that might take place within the upcoming one to two weeks. This feature is deployed with the use of a huge language model provided via an external application programming interface. The design is based on the emerging evidence that conversational systems can facilitate emotional reflection and self awareness provided that they are applied with appropriate guardedness.The response of the person to all the forty two items of DASS,the occurrence of the intensity of symptoms is summarized in the form of a short natural language,the probability of risking the highest class of the machine learning model,mental health information that has been pre selected and retrieved and incorporated as context within a small internal knowledge base.These were the inputs that were used for the language model

3.5.1 Question Generation with Personalisation

On this information, the language model produces precisely five open ended questions of reflection which is used to investigate the recent emotional fluctuation, day to day functioning

, present situation or stress causing situations , the anticipations of the future and finally the sources of support or strain.The questions are worded in colloquial language and purposely not in clinical language or diagnostic terminology or definite mentioning of risk scores. This aligns with the suggestions of a safe and supportive digital mental health tool, in which the emphasis is placed on the reflection, rather than diagnosis.

3.5.2 Short Term Emotional Tendency Estimation

The same language model generates a qualitative summary of the possible short term emotional trajectories after the user has typed answers to the five questions and it approximated 3 chances, Likely the mood and stress levels will be generally similar,The likelihood of the individual feeling significantly more positive in general,The likelihood that the individual is more significantly worsened or overloaded.These probabilities are bound to be in the range between zero and one and to add to be roughly equal to one. The model then creates a brief explanation in common language to explain what these figures are all about. It is an output that is placed as a reflective and speculative account rather than a clinical prediction and treatment recommendation.

3.6 System Architecture

The entire system is a prototype web-based system implementation and is made of three major parts

Machine Learning Layer

The machine learning model analyses the surveydata an inturn calculates the dass sub scales to determine the final score,reads the survey data and calculates DASS sub scale and total scores. ‘it also forms three class risk label in accordance to the DASS score.It runs the theme of the sequence of the preprocessing on numerical and categorical variables. Many multi class classifiers are trained and tested.

Language Based Reasoning Layer

Accepts the questionnaire answers and risk estimate of the chosen individual. Bases on the internalised knowledge context along with the language model to produce the personalised reflective questions. Takes free text answers of the user and generates a short term emotional prognosis in the form of three probabilities and a description.

Figure 4: The Flow of the Streamlit User interface

Interactive Front-End

An intereactive web interface was deployed as a web interface on rapid application development framework which Enables the user to browse through test set participants and shows the current risk estimates, DASS scores, and all the response sets in a questionnaire and also displays the language model and has text boxes that the user can type in the answers and finally it displays the resultant short-term probability breakdown and description in a presentable format.The machine learning model will provide numbers, representing the probability, which is converted to natural language by the reasoning layer to produce an intuitively understandable description of what might happen in the short term with respect to emotions.

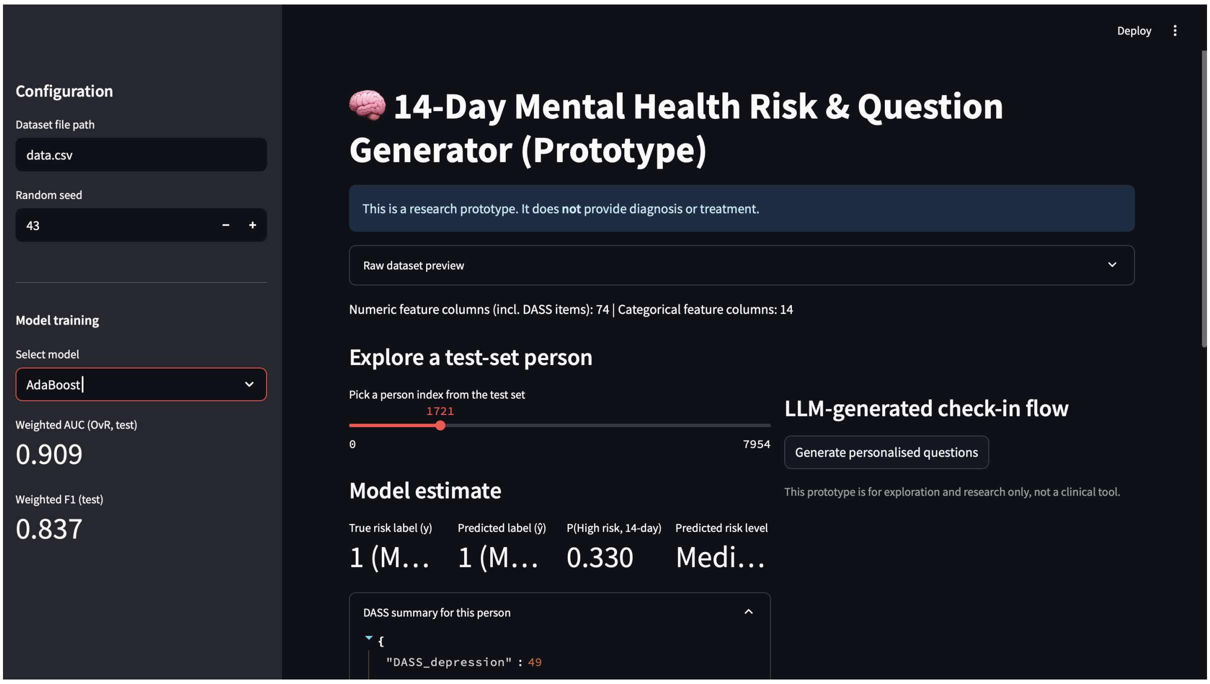

Figure 5: Interactive front end layer

The user interaction narrative shows the entire process of marketplace usage of the Streamlit application interface. As a part of Step 1, the user opens the Streamlit application and chooses a record of the test data to analyze. It then moves on to Step 2 that shows the machine learning findings containing the subscale scores of Depression Anxiety Stress Scales, the probability of risk that is currently present, based on the logistic regression model, and the risk classification that is predicted.When the user is at Step 3, the large language model component is actively involved, where he/she clicks on the "Generate personalised questions" button and the question generation process ensues. This culminates into Step 4 wherein the system delivers five personalized, open ended reflective questions that are specific to the risk profile and questionnaire answers of a particular user. Free text responses are then entered by the user directly into the application interface.And lastly, Step 5, the user responses are processed by the large language model summary step then the seven to fourteen day outlook is presented to the user. This perspective gives the probability estimates of the three possible mood trajectories, that is, to stay the same, improve or decline. This last action serves as the final cycle in the interaction that offers the user not only instant reflective services but also prospective predictive data about their emotional state of well being.

3.7 Ethical Considerations

The system is not of clinical research prototype but is a diagnostic/treatment tool and is not designed to substitute professional assessment and treatment; instead, the system is meant to help the individual reflect and become familiar with emotional patterns at an early stage.All of these considerations are meant to guarantee that the hybrid system serves as an auxiliary and exploratory resource in a research setting in consideration of the sensitivity of mental health data and constraints of computational forecasting.

4 Design Specification

The proposed system will be a hybrid predictive reflective system, which integrates supervised machine learning methods with large language model reasoning. The essence of this architecture is to determine the current psychological risk level of an individual and give a short term emotional perspective that can be understood. The architecture is built based on a number of interrelated modules, such as data ingestion, psychological scoring, feature engineering, multi class classification modelling and generative reasoning layer.The data is analyzed with the DASS-42 scoring system that generates valid knowledge of Depression, Anxiety, and Stress. These subscales are added to one scale DASStotal, which is the total level of emotional distress of the individual participants. Another element proposed in this project is a quantile based triage algorithm that will transform the continuous DASStotal values into three equally balanced risk groups: low risk, medium risk, and high risk. This group of them is the target variable of the machine learning models.The construction of feature engineering consists of two categories, namely numerical features (survey timing, personality indicators, vocabulary recognition scores, demographic measures, and raw item responses), and categorical features (demographic and contextual variables).

4.1 Pychological scoring

The dataset originates from the structured DASS-42 questionnaire, complemented with demographic indicators, personality traits, vocabulary recognition metrics, and timing metadata. Using the scoring rules described by Lovibond and Lovibond (1995), the Depression, Anxiety, and Stress subscales are computed and then aggregated into a composite index referred to as DASS-total. This score represents an individual's overall emotional distress level.To prepare the data for supervised classification, a quantile-based triage algorithm converts continuous DASS-total values into three statistically balanced categories representing low-, medium- and high-risk groups. This strategy avoids dominant-class bias and ensures even distribution across the emotional-risk spectrum, resulting in a stable predictive task.

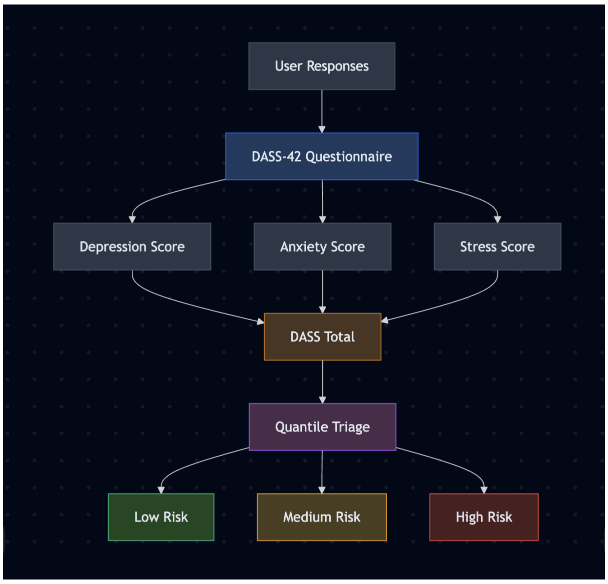

Figure 6: Psychological scoring workflow

The figure shows how raw responses by the participants on the DASS-42 questionnaire are converted into validated Depression, Anxiety and Stress subscale scores. These subscales are combined into one of the distress indicators (DASS total), which is, in turn, subjected to a triage scheme based on the quantile and the resulting risk (low, medium, high) is used in further machine-learning modelling.

4.2 Feature Engineering architechture

The system incorporates a structured feature-engineering pipeline built using a ColumnTransformer design. This enables numerical and categorical inputs to pass through dedicated transformation paths, ensuring that each variable type receives an appropriate mathematical treatment.The numerical component consists of DASS subscale values, vocabulary recognition scores, demographic numbers, timing information, and raw item responses. The categorical component includes demographic labels, personality attributes, and contextual variables. Numerical features undergo median imputation to address outliers and inconsistencies, followed by standard scaling to normalise the feature space. Categorical variables pass through most-frequent imputation and one-hot encoding to ensure compatibility with all downstream models.This modular preprocessing ensures reproducibility, prevents data leakage, and standardises the feature representations used across all classifiers.

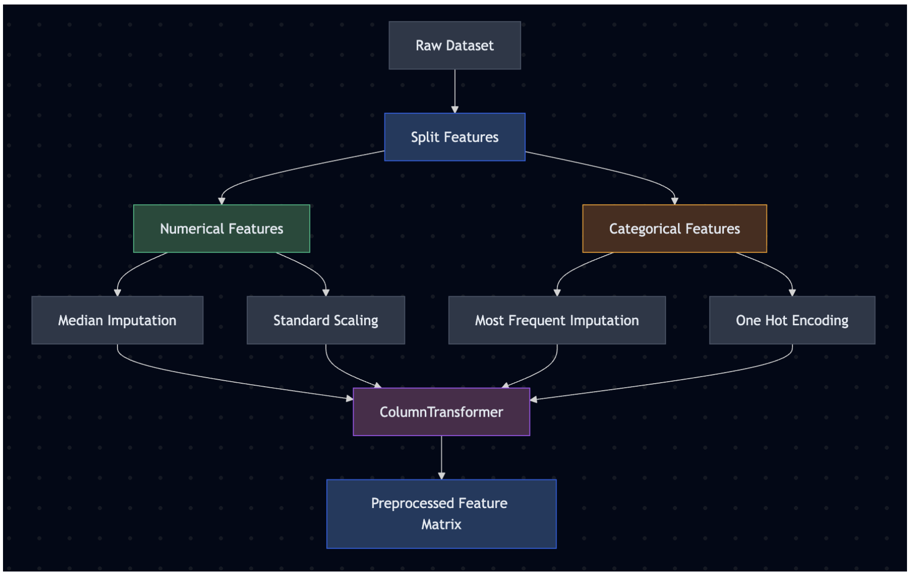

Figure 7: Feature engineering and preprocessing pipeline

The workflow of feature-engineering is shown in the diagram, where the first stage of work is the splitting of raw data into numerical and categorical parts. All the feature types are subject to the proper preprocessing pathway median imputation and scaling of numerical variables, and imputation using one-hot encoding of categorical variables. These transformations are combined in a ColumnTransformer that results in a completed processed feature matrix that can be analyzed to be trained into a model.

5 Implementation

This implementation phase resulted in a full-fledged hybrid system that includes a monitored machine-learning pipeline and a large language model reasoning component and makes the overall functionality of the hybrid available as an interactive web interface. The basic dataset was optimized and enriched with the calculated DASS subscale scores (Depression, Anxiety, Stress), a DASS-total summary score and calculated three-level risk label (Low, Medium, High) based on sample quantiles. This set of derived variables along with demographics, personality, vocabulary recognition and timing metadata comprise the canonical dataset of the predictive pipeline.The final modelling inputs were the results of data cleaning and preprocessing which were applied in a deterministic fashion. Median values were used to impute numeric values in missing cases, and were brought to zero mean and unit variance to make the estimates of the coefficients and distance-based algorithms stable. The most common category was used to impute categorical variables and to be encoded using one-hot representations in a way that categorical information could be fed to the classifiers. An orchestration was used to provide type preserving transformations such that each feature was sent to the correct imputation and scaling path such that the pipeline could be serialised and then used in the same mode during training and deployment.

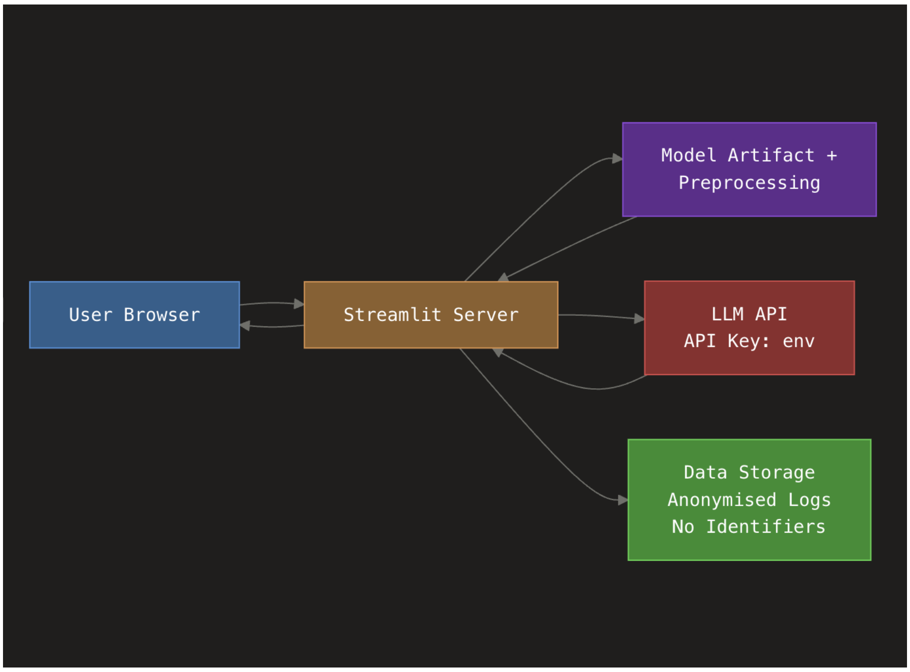

Figure 8: System deployment workflow

The figure explains the interaction of the user with the Streamlit server which loads the stored ML model and preprocessing pipe, securely interacts with the external LLM API using keys stored in the environment, and optionaly logs anonymous data to storage. The complete flow of data running during the run time is displayed by this layout and it provides privacy as no personal identifiers are stored.

5.1 Machine learning component

The predictive modelling part measured a collection of existing supervised classifiers which had been chosen due to their robustness, interpretability and comparative benchmarking. The implemented models in this paper were multinomial logistic regression, random forest, gradient boosting, AdaBoost, support vector classifier, K-nearest neighbours and XGBoost. All these algorithms used the transformed feature vectors generated by the preprocessing pipeline making inputs consistent at the modelling environment.To enhance transparency and also to find out consistent parameter combinations, a systematic hyperparameter search was conducted by use of grid search and random search methods. In the case of logistic regression, optimisation was taken into account L2 regularisation with penalty weight values of C = 0.1, 1, 10 and applied the LBFGS solver. The exploration that was conducted in the case of random forests was based on the variation of the number of estimators between 100 and 500, the maximum tree depth of 5 to 50, and the minimum sample split of 2 to 10. The models of gradient boosting were optimized with the trial of learning rates 0.01 and 0.1 and estimators 50 to 200. In the case of XGBoost, the values of the maximum depth have been changed to 3-7; learning rates of 0.05-0.3; and subsampling between 0.7 and 1:1.

The argument behind the choice of logistic regression as the ultimate predictive model was based on its interpretability, the fact that it produces calibrated probabilities, and it is stable in activities that are being characterised by the imbalance of classes. In spite of the fact that gradient boosting and XGBoost proved to be comparable in terms of accuracy, logistic regression offered a more straightforward mapping of feature contribution to the risk levels predicted, and its predictive compliance was more in line with the ethical and methodological concerns of a system based on mental health focus. However, the competing models were included all in experimentation and comparison.

5.2 Large Language model Component

The LLM reasoning module is an addition to the numerical classifier that provides reflexivity insights. It takes structured inputs in the form of DASS subscale scores, emotionally-at-risk or emotionally-not-at-risk predicted probabilities, a natural-language summary of the symptoms and free-text user-provided reflections. All these inputs are incorporated into a well-considered prompt that is written in non-diagnostic language that is safe and in line with psychological practices.The LLM does three fundamental reasoning functions. It also produces individualized reflective questions that are specific to the risk profile that is predicted by the individual. Second, it reads the free-text answers of the user, uses sentiment and context analysis to learn trends in the emotional experience. Third, it develops a qualitative perspective in how the emotional state of the individual will develop within the following 7-14 days in either supportive and non-clinical terms.It is a method that brings efficiency to numbers and brings people to the interpretability of natural language to allow the system to assist in reflective emotional awareness but not give clinical advice.

5.3 Component deployment diagram

The deployment diagram shows the interaction of the user with the Streamlit application that loads the preprocessing pipeline and machine-learning model, sends structured prompts to the external LLM API and logs only anonymised logs without any personal identifiers. The layout emphasizes the modular data movement and provides privacy-preserving communication between all the system components.

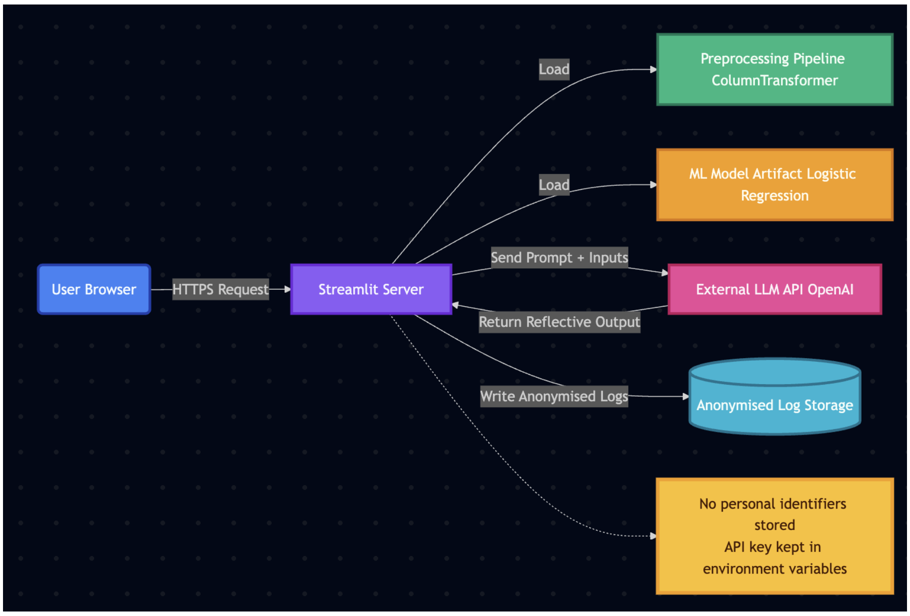

Figure 9: Component deployment

6 Evaluation

The test of the suggested hybrid predictive reflective system was achieved by means of a set of systematic experiments. The created experiments aimed to evaluate (1) the predictive accuracy of the machine learning part, (2) how well the model generated probabilities can be predicted and interpreted and (3) the behaviour of the large language model reasoning layer. Compared to each other, these assessments will answer the research question by establishing whether the system can provide meaningful emotional risk differentiation and consistent short term reflective assessments.

Experiment 1: Multiclass Classifier Performance

The initial study reviewed the effectiveness of the monitored machine learning models that were trained to estimate the low, medium, and high emotional risk levels. Five models were compared to logistic regression, random forest, gradient boosting, AdaBoost and XGBoost. The assessment was conducted on a stratified hold out test set based on the weighted ROC-AUC and weighted F1-score, to consider minimal differences in the distribution of the classes.The findings showed that gradient boosting had the best weighted F1-score of 0.8433 and weighted ROC-AUC of 0.9569. XGBoost came right after this performance with a weighted F1-score of 0.8399 and a weighted ROC-AUC of 0.9575. Competitively, logistic regression obtained a weighted F1-score of 0.8380 and weighted ROC-AUC of 0.9561.

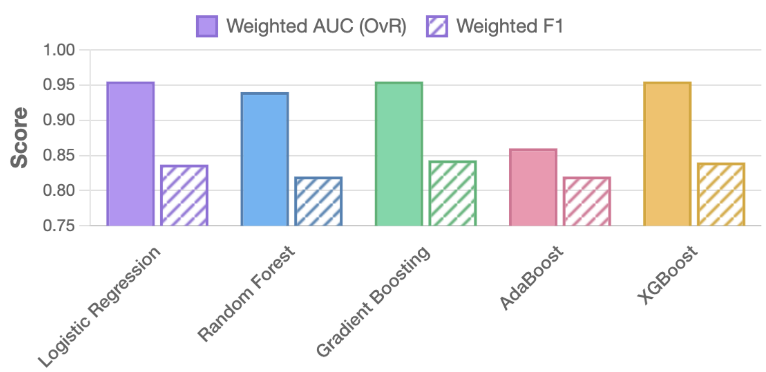

Figure 10: Model results

Random forest and AdaBoost were showing relatively poorer results, indicating that more complex ensembles and less informative boosting strategies did not fit the structural dimensions of the psychological survey data.Bar charts, histograms, and boxplots were used to confirm that the dataset had a large overlap between medium-risk and high risk levels. This can be used to understand why the separability between these categories is lower in model predictions. Further confusion matrices showed that the most frequent cases of misclassification occurred at the medium high boundary, which shows that, due to the subjective and continuous nature of emotional distress, these two groups are inherently hard to differentiate.The experiment showed that the models can significantly discriminate between broad emotional risk categories despite the fact that clinical accuracy is not possible based on this data.

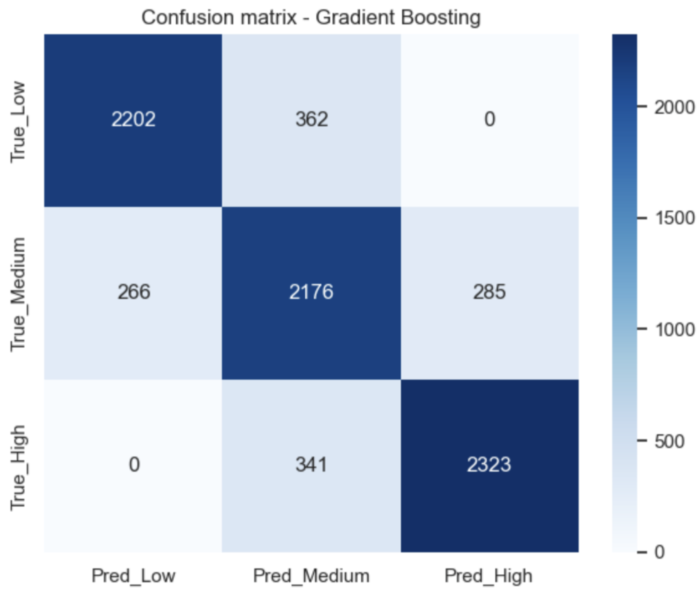

Figure 11: Confusion Matrix for Gradient Boosting

The matrix indicates that high accuracy rates are achieved in the low risk cases and the majority of the misclassifications are between the medium and high risk groups as the self reported distress scores are continuous and overlapping.

Experiment 2: Model Probabilities Stability and Calibration

The second experiment compared the interpretability, smoothness and calibration of predicted risk probability of the various classifier. This analysis was needed since probability estimates as opposed to discrete class labels are the input of reasoning layer of the system.The most consistent and monotonic transitions of the probability between the low, medium and high risk classes were the result of a Logistic regression. This consistency is what makes it especially appropriate in situations where interpretability and transparency are the key factors. Gradient boosting and XGBoost produced more confident and sharper decision boundaries which boosted accuracy but created increased variance between edges of classes.There was a zone of decreased confidence range between the medium and high risk brackets in all the models. This behaviour confirms the result of Experiment 1 and can be once more credited to the incessant, as opposed to discrete, quality of psychological distress. Nevertheless, this ambiguity, even though it was inherent, was sufficiently mitigated in the sense that the models were capable of providing meaningful downstream reasoning.This experiment approved the selection of the logistic regression as the main model of the prototype, which balances accuracy and interpretability and has a stable behaviour with imbalanced classes.

Experiment 3: Characterization of the Large Language Model Reflective Reasoning Layer

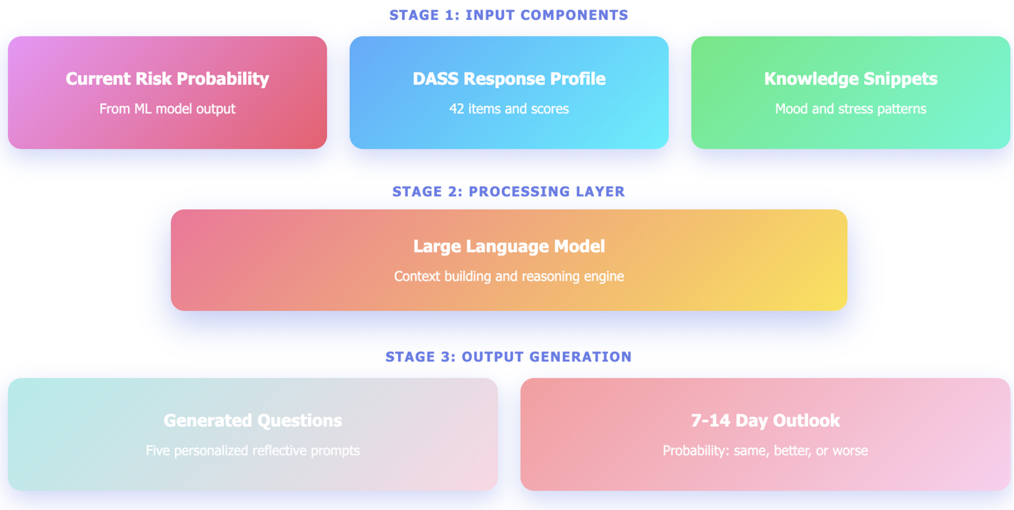

Figure 12: Pipeline for the Large Language model

The qualitative performance of the large language model component was tested in the third experiment. The module produces bespoke reflective queries and a low range emotional perspective in regards to the forecasted threat likelihood and a natural language overview of responses on DASS questions.The findings showed that the model was consistent in producing coherent and non clinical reflective questions, which complied with the psychological safety standards. The questions were non diagnostic and rather inviting to emotional reflection in a language that is easy to understand, which corresponds with the ethical principles of the project.In the case of free text simulated responses, the model would produce structured short term emotional predictions that are broken down into three likely trajectories: remaining steady, improving or getting worse in the next 7–14 days. These synoptics had been prepared in encouraging and moderate language, without taking a stand, or making clinical assertions. Even though the system does not execute clinical reasoning, its outputs exhibited internal thematic coherence and practical appropriateness in the reflective application.This experiment validated the fact that the reasoning layer was able to conduct a successful transformation of the outputs of the numerical model into reflective insights that are understandable to human beings, which fulfilled the second objective of the research.

6.1 Discussion

This research evidence suggests that the hybrid predictive reflective system offers a practical and theoretically encouraging strategy of generating mental health insights in the short term. The machine learning models had moderate performance as measured by classification with the highest performing models having a weighted F1 score between 0.84. These values demonstrate that the model is capable of effectively differentiating the broad categories of emotional risk, but not with clinical accuracy. This is a level of performance in line with other previous studies that used analogous questionnaire based datasets, as the high level of symptom overlap and subjective reporting tends to inhibit accuracy.The results of the classifiers indicate a common issue with mental health prediction studies the gray zone between moderate and severe distress.This methodological limitation should be considered, and it implies the impact on the interpretability and on generalisability. Nevertheless, logistic regression was a useful option in the presence of this drawback because it is stable even when there is a class imbalance and is interpretable, which is essential when implementing it in a sensitive environment ethically.The language model part to be an effective and supportive problem analysing tool, producing sensible questions and summarising probable emotional patterns in the next one to two weeks. These were not the diagnostic outputs, of which it is appropriate that situates safe mental health technology guidelines. Nevertheless, big language models cannot be based on clinical reasoning, but through linguistic pattern matching, hence their interpretations should be approached with care. The sensitivity of their responses to the way they are phrased by the user makes them variable, and uncontrollable to a considerable extent. Irrespective of these shortcomings, the reflective reasoning layer is beneficial in the sense that it transforms the numerical prediction experience to a personalised experience.Combined, the experiments prove that the system is successful in attaining its conceptual goals but is inappropriate to clinical or high stakes use. This lack of longitudinal data and the artificial stratification of the risk groups together with the subjective character of the inputs restrict the level of insight that can be provided by the model. However, the system is efficient as a research prototype that is aimed to assist in the initial emotional and exploration usage.

7 Conclusion and Future work

In this project, a hybrid system combining supervised machine learning with a large language model was presented and uses it to offer personalised and understandable short term mental health information. The primary predictive model was chosen to be logistic regression because this model displayed stable behaviour, it was interpretable and competitive compared to more complex alternatives. Its probability estimates were entered into the language based reasoning part that generated individualised reflective questions and a brief story about how the emotional state of one person may be transformed amongst the next one to two weeks.The system provides two types of output one quantifiable measure of current emotional risk, and a qualitative description of a short term emotional tendency. This stratified technique offers more available and helpful feedback compared to numerical predictions only. Despite the limitations in the system presented by the unavailability of longitudinal data, the use of self reported questionnaire data, and theoretic weakness of the generative models, the results show that the predictive modelling and reflective reasoning types of the system can deliver meaningful user experiences in digital mental health tools.The incorporation of behavioural signals, longitudinal tracking, better model calibration and structured evaluation with actual users could be added to the further development. These improvements would create a basis of more flexible, moral and personalised mental health technologies. The project shows that combining structured prediction with facilitating reflection is feasible and has potential even in research prototype prototypes.

References

Aledavood, T. et al. (2025). Multimodal digital phenotyping study in patients with major depressive episodes and healthy controls (mobile monitoring of mood): Observational longitudinal study, *JMIR Mental Health* **12**: e63622. Spring.

Beames, J. et al. (2024). Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12–25 years): A scoping review, *SSRN Electronic Journal* .

Bhaganagare, S., Chavan, S., Gavali, S. and Godase, V. (2025). Mood mate: A review of multimodal ai in real-time mental health monitoring, *Unpublished manuscript / Review* . Preprint.

Bufano, P., Laurino, M., Said, S., Tognetti, A. and Menicucci, D. (2023). Digital phenotyping for monitoring mental disorders: Systematic review, *Journal of Medical Internet Research* **25**: e46778.

Cai, Y., Su, X., Yao, D., Han, R., Gao, N. and Zhang, H. (2025). Design and challenges of mental health assessment tools based on natural language interaction, *arXiv preprint arXiv:2510.18158* .

Choi, A., Ooi, . and Lottridge, D. (2024). Digital phenotyping for stress, anxiety and mild depression: A systematic literature review (preprint), *JMIR mHealth and uHealth* **12**: e40689–e40689.

Cruz-Gonzalez, P. et al. (2025). Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications, *Psychological Medicine* **55**.

Daga, V. (2025). Ai for early detection on mental health via passive mobile health sensing, `https://www.ijrrr.com/specialissues2-2025/ijrrr-Special-Issue-2-2025-paper21.pdf`. Accessed: 2025-12-09.

Fatouros, P., Tsirmpas, C., Andrikopoulos, D., Kaplow, S., Kontoangelos, K. and Papageorgiou, C. (2025). Randomized controlled study of a digital data driven intervention for depressive and generalized anxiety symptoms, *npj Digital Medicine* **8**(1).

Jacobucci, C. and Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement, *Perspectives on Psychological Science* **15**(3): 809–816.

Jung, H. W. et al. (2025). Key features of digital phenotyping for monitoring mental disorders: A systematic review (preprint), *Journal of Medical Internet Research* .

Lovibond, S. H. and Lovibond, P. F. (1995). *Manual for the Depression Anxiety & Stress Scales*, 2 edn, Psychology Foundation, Sydney, Australia.

Lundberg, S. M. et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, *Nature Biomedical Engineering* **2**: 749–760.

Luque, A. et al. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix, *Pattern Recognition* **91**: 216–231.

Nemesure, M. et al. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence, *Scientific Reports* **11**(1).

Shvetcov, A. et al. (2024). Passive sensing data predicts stress in university students: a supervised machine learning method for digital phenotyping, *Frontiers in Psychiatry* **15**.

Tafti, T. et al. (2021). Predicting mental health from followed accounts on twitter, *Journal of Medical Internet Research* **23**(4): e25987.

Terhorst, Y., Messner, E.-M., Opoku Asare, K., Montag, C., Kannen, C. and Baumeister, H. (2025). Investigating smartphone-based sensing features for depression severity prediction: Observation study, *Journal of Medical Internet Research* **27**: e55308.

Torous, J. et al. (2025). The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality, *World Psychiatry* **24**(2): 156–174.

Zhang, Y. et al. (2025). Large-scale digital phenotyping: Identifying depression and anxiety indicators in a general UK population with over 10,000 participants, *Journal of Affective Disorders* .