BCSE497J Project-I

DETETECTION OF EARLY MENTAL HEALTH DISORDERS USING EXPLAINABLE ML APPROACH

22BCE0490 ABHAY SINGH

22BCE2573 MANAS JOSHI

22BCE2588 YAGNIT MAHAJAN

Under the Supervision of

Dr. Yoganand S

Assistant Professor Sr. Grade 1

School of Computer Science and Engineering (SCOPE)

B.Tech.

in

Computer Science and Engineering

School of Computer Science and Engineering

September 2025

ABSTRACT

Mental health challenges such as anxiety, depression, and addiction often go unnoticed until they escalate into severe conditions, making early detection crucial for effective intervention. This project focuses on building an AI-powered system that can analyze Reddit posts to identify early signals of these conditions through advanced natural language processing techniques. Since Reddit serves as an open platform where individuals frequently share personal experiences and struggles without the fear of social stigma, it provides a rich and authentic source of unstructured data for mental health research. By gathering fresh posts through sophisticated web scraping methodologies, the project aims to capture real-time discussions around mental health topics, ensuring that our model learns from diverse and constantly evolving language patterns. This approach emphasizes both technical relevance and human empathy in tackling the pressing issue of mental health awareness in our digital age.

To achieve robust classification capabilities, a transformer-based model like BERT (Bidirectional Encoder Representations from Transformers) will be fine-tuned specifically for multi-label classification tasks. Unlike traditional single-label models, this approach acknowledges the complex reality that people may experience multiple overlapping conditions simultaneously-for example, anxiety coupled with insomnia, or depression intertwined with substance abuse patterns. The system architecture will incorporate advanced preprocessing techniques, including text normalization, tokenization, and semantic analysis to handle the informal nature of social media language. Furthermore, the model will be trained using carefully curated datasets with proper class balancing techniques to avoid bias toward more commonly reported conditions. This comprehensive approach ensures that our system can accurately identify subtle linguistic markers that may indicate emerging mental health concerns.

The system will not only classify posts into relevant categories but also provide complete transparency through SHAP (SHapley Additive exPlanations), a cutting-edge explainable AI methodology. SHAP highlights which specific words, phrases, or linguistic patterns influenced the model's prediction, ensuring that users, researchers, and healthcare professionals can trust and thoroughly understand the system's reasoning process. This interpretability component is crucial for building confidence among medical practitioners who might integrate such tools into their diagnostic workflows. Additionally, the explainability features will help identify potential biases in the model's decision-making process, allowing for continuous improvement and ethical refinement of the algorithm.

The comprehensive insights generated by our system will be presented through an interactive React-based dashboard, meticulously designed to visualize emerging mental health trends, risk level distributions, and user-friendly explanations of model predictions. This intuitive platform will empower diverse stakeholders-including healthcare professionals, mental health researchers, policy makers, and even general users-to explore complex patterns, track condition prevalence over time, and gain actionable awareness about community mental health status. By seamlessly combining advanced AI/ML techniques, robust explainability frameworks, and thoughtful user experience design, this project contributes directly to achieving the United Nations Sustainable Development Goal 3: Good Health and Well-being. Ultimately, it aspires to spark early awareness, empower timely intervention strategies, and significantly reduce the persistent stigma surrounding mental health discussions, demonstrating how modern technology can serve not just as a predictive tool but also as a meaningful bridge connecting data science with human empathy and community support systems.

TABLE OF CONTENTS

S. No.	Contents	Page No.
	ABSTRACT	**2**
1.	**INTRODUCTION**	**4**
	1.1 BACKGROUD	
	1.2 MOTIVATIONS	
	1.3 SCOPE OF THE PROJECT	
2.	**PROJECT DESCRIPTION AND GOALS**	**5-7**
	2.1 LITERATURE REVIEW	
	2.2 GAPS IDENTIFIED	
	2.3 OBJECTIVES	
	2.4 PROBLEM STATEMENT	
	2.5 PROJECT PLAN	
3.	**REQUIREMENT ANALYSIS**	**8**
	3.1 FUNCTIONAL REQUIREMENTS	
	3.2 NON-FUNCTIONAL REQUIREMENTS	
	3.3 SOFTWARE REQUIREMENTS	
	3.4 DATASET REQUIREMENTS	
	3.5 USER REQUIREMENTS	
4.	**SYSTEM DESIGN AND PROJECT PLANNING**	**9-13**
	4.1 PROJECT PLANNING	
	4.2 HIGH-LEVEL DESIGN (HLD)	
	4.3 LOW-LEVEL DESIGN (LLD)	
5.	**REFERENCES**	**14**

1. INTRODUCTION

Mental health is a critical aspect of overall well-being, yet it remains one of the most under-addressed areas of healthcare worldwide. Conditions like anxiety, depression, and addiction often manifest subtly, leaving them unnoticed until they become severe and harder to manage. At the same time, online platforms have emerged as spaces where people openly express their emotions and struggles, particularly on Reddit, where anonymity encourages honest sharing. With advancements in artificial intelligence and natural language processing, these vast discussions can now be analyzed to detect early signs of mental health concerns. This project builds on that opportunity, aiming to harness AI not only for detection but also for raising awareness and promoting early interventions.

1.1 BACKGROUND

Traditional approaches to diagnosing or monitoring mental health rely on self-report surveys, clinical interviews, and psychological assessments. While effective in structured settings, they are often limited by accessibility, affordability, and social stigma. Meanwhile, social media platforms provide unfiltered, large-scale data that reflects real-time human experiences and emotional expressions. Reddit is particularly valuable because its community-driven structure fosters detailed conversations around sensitive topics, including mental health. At the same time, the emergence of transformer-based models like BERT has transformed the ability of machines to understand nuanced language, including slang, abbreviations, and emotional cues common in online discussions. This technological and social context forms the foundation for the proposed project.

1.2 MOTIVATIONS

The motivation for this project is rooted in both societal impact and technological advancement. On a human level, there is an urgent need to address the growing burden of mental health issues, particularly in young adults who frequently engage with online platforms. By detecting early signals from their everyday conversations, this project can help raise awareness before conditions worsen. On a technological level, the project seeks to explore the potential of advanced AI models like BERT in real-world healthcare applications, while ensuring transparency through explainable AI methods such as SHAP. Together, these motivations underline the project's commitment to developing an AI system that is not only powerful but also ethical and socially meaningful.

1.3 SCOPE OF THE PROJECT

The scope of this project is centred around analysing Reddit posts to detect multiple mental health conditions, including but not limited to anxiety, depression, and addiction. The methodology involves collecting data through web scraping, preprocessing text, and fine-tuning a BERT-based model for multi-label classification. SHAP will be employed to provide interpretability, ensuring that predictions are explainable and trustworthy. Finally, the results will be visualized through a React-based dashboard, offering stakeholders—such as healthcare professionals, researchers, and the general public—an accessible way to explore patterns and trends. While the system is not intended to act as a diagnostic tool, it serves as an awareness and research platform that directly contributes to the United Nations Sustainable Development Goal 3: Good Health and Well-being.

2. PROJECT DESCRIPTION AND GOALS

This section outlines the project's foundation, reviewing existing literature, identifying research gaps, defining objectives and problem statement, and presenting a structured plan to achieve its goals.

2.1 LITERATURE REVIEW

The application of artificial intelligence for mental health detection has evolved significantly, with recent research emphasizing explainable AI approaches and multi-modal learning. Zogan et al. (2020) introduced the Multi-Modal Depression Detection with Hierarchical Attention Network (MDHAN), combining deep learning with interpretability through two-level attention mechanisms at tweet and word levels. Their work demonstrated that explainable models could outperform traditional baselines while providing transparent reasoning for depression detection. However, this approach focused primarily on binary classification rather than comprehensive multi-class scenarios.

Recent advances have explored ensemble transformer-based models integrating personality and sentiment patterns for enhanced mental health detection. Kerz et al. (2023) conducted systematic investigations of explainable AI techniques, evaluating BiLSTM models with human-interpretable features across five mental health conditions (ADHD, anxiety, bipolar disorder, depression, and psychological stress). Their research revealed that personality-infused models achieved highest accuracy for anxiety (73.40%) and bipolar conditions (73.23%), while demonstrating the effectiveness of LIME and attention-based explanations. Similarly, contemporary studies have developed ensemble methods with iterative majority voting for efficient psychiatric disorder detection, emphasizing computational efficiency balanced with accuracy for real-world deployment.

Despite these advances, significant research gaps persist. Most existing studies focus on binary classification or limited disorder categories (typically 2-5 conditions), while comprehensive multi-class frameworks addressing 10+ mental health disorders simultaneously remain largely unexplored. Additionally, while explainable AI techniques are increasingly integrated, few studies embed explainability as a core architectural component rather than post-hoc additions. This creates opportunities for developing comprehensive, explainable systems that can detect multiple mental health conditions while providing transparent, clinically-relevant reasoning for healthcare applications.

2.2 GAPS IDENTIFIED

Based on comprehensive analysis of existing literature, several critical gaps emerge that restrict the practical applicability and clinical utility of current mental health detection systems. A fundamental limitation across studies is the over-reliance on binary classification frameworks that fail to reflect mental health complexity. Models such as MDHAN (Zogan et al., 2020), ensemble transformer-based approaches (2024), social media mental health studies (2023-2024), psychiatric disorder detection models (2024), and systematic XAI evaluations (Kerz et al., 2023) focus predominantly on distinguishing between presence and absence of specific conditions, overlooking the reality that mental health conditions often co-occur simultaneously. Additionally, explainability mechanisms including attention visualizations target technical audiences while providing limited value to healthcare professionals who require clear, actionable insights for clinical decision-making.

The complexity of multi-modal architectures presents another significant challenge across these studies. While ensemble transformer-based models integrating personality traits and sentiment patterns demonstrate strong accuracy, these approaches demand substantial computational resources and intricate feature engineering pipelines, substantially reducing feasibility for real-world clinical deployment where efficiency and reliability are paramount. Current research also exhibits notably narrow focus on specific disorders, with most studies concentrating heavily on depression while disorders such as anxiety, ADHD, autism, bipolar disorder, OCD, and PTSD remain significantly underexplored.

The absence of real-time monitoring and deployment strategies further limits clinical value across all reviewed approaches. Continuous tracking and early intervention capabilities are critical for effective mental health care, yet research remains confined to static analyses without addressing temporal dynamics or deployment feasibility. These limitations collectively underscore a critical research opportunity: developing a computationally efficient, explainable AI framework capable of detecting and differentiating across a broad spectrum of mental health disorders while providing transparent, clinically relevant explanations accessible to diverse end-users.

2.3 OBJECTIVES

The project is guided by the objective of converting unstructured social media narratives into meaningful insights that support the early detection of mental health conditions using advanced AI techniques. The primary aim is to build a comprehensive explainable AI system for multi-class mental health disorder detection from social media text, providing transparent and clinically relevant insights that can be easily understood by healthcare professionals and researchers.

The first specific objective is to design robust classification models capable of distinguishing between multiple mental health categories. The system will focus on detecting more than 11 disorders including anxiety, depression, ADHD, autism, bipolar disorder, borderline personality disorder (BPD), OCD, PTSD, psychosis, addiction, and suicide ideation. To achieve this, it will combine traditional feature engineering approaches such as TF-IDF and sentiment analysis with advanced transformer-based architectures like BERT and RoBERTa. Class weighting techniques will also be applied to address the imbalance that is common in mental health datasets.

The second objective is to integrate explainable AI methods such as SHAP (SHapley Additive exPlanations) and attention visualization techniques. This will make the decision-making process of the models fully transparent. The system will highlight the words, phrases, and linguistic patterns that contribute most to predictions, offering explanations that healthcare professionals can validate against clinical knowledge and diagnostic criteria.

The third objective is to develop an intuitive web-based interface where users can input text and receive real-time predictions along with clear explanations. The interface will display prediction confidence scores, feature importance rankings, and highlighted text segments. This design will ensure that the system can be used by healthcare professionals, researchers, and advocacy groups without requiring technical expertise in machine learning.

The final objective is to carry out a rigorous performance evaluation using stratified cross-validation methods, comparison with baseline models, and qualitative analysis of interpretability outputs. The evaluation will measure accuracy, precision, recall, and F1-scores across all categories, with special attention to minority classes. It will also include expert reviews and user feedback to validate the clinical relevance and practical applicability of the explainability features in real-world healthcare settings.

2.4 PROBLEM STATEMENT

Despite significant progress in applying artificial intelligence for mental health detection, existing systems remain constrained by binary classification frameworks, limited disorder coverage, and a lack of clinically relevant explainability. Most approaches focus narrowly on depression or a small set of disorders, often using complex multi-modal models that are difficult to deploy in real-world healthcare environments. Furthermore, current explainability methods are primarily technical in nature, providing little actionable value for healthcare professionals and patients.

There is a pressing need for a comprehensive, explainable AI system capable of detecting a wide range of mental health disorders from unstructured social media narratives. Such a system must combine advanced classification techniques with transparent interpretability features, ensuring that predictions are not only accurate but also understandable and clinically meaningful. Addressing this problem would support early detection, timely intervention, and improved trust in AI-driven mental health tools.

2.5 PROBLEM PLAN

The project will begin with data-driven foundations by leveraging a curated dataset of 12,471 Reddit posts collected from 11 mental health–focused communities. A systematic preprocessing pipeline involving deduplication, spam removal, and text cleaning has ensured data quality, with the final dataset offering diversity across multiple categories, balanced representation, and natural self-expressive text for modeling. This dataset serves as the basis for developing robust multi-class classification systems and explainability techniques.

Model development will combine traditional feature engineering methods such as TF-IDF and sentiment analysis with state-of-the-art deep learning approaches, particularly transformer-based architectures like BERT. Machine learning models including Logistic Regression, SVM, and Random Forest will also be implemented for comparative analysis. Class weighting strategies will be employed to address any residual imbalance across categories. To enhance transparency, explainable AI techniques such as SHAP, LIME, and attention-based visualization will be integrated, allowing identification of key words and linguistic patterns responsible for predictions.

Finally, the project will deliver an intuitive web-based application built with React.js for the frontend and FastAPI for backend model services. The application will provide real-time predictions, confidence scores, and interpretable explanations through interactive visualizations. Performance evaluation will be carried out using stratified cross-validation, comparative benchmarking against baselines, and expert feedback on the quality of explanations, ensuring both technical robustness and clinical relevance.

3. REQUIREMENT ANALYSIS

3.1 FUNCTIONAL REQUIREMENTS

The system shall classify social media posts into 11 mental health categories, including anxiety, depression, ADHD, autism, bipolar disorder, borderline personality disorder (BPD), OCD, PTSD, psychosis, addiction, and suicide ideation. Predictions should be accurate, interpretable, and generated in real-time. The system shall integrate explainable AI techniques, including SHAP and attention-based visualizations, to highlight key words and phrases influencing each classification.

The system shall provide an interactive web interface enabling users to input text, view prediction confidence scores, and explore interpretable explanations. Additionally, it shall support automated model training, evaluation with standard metrics (accuracy, precision, recall, F1-score), and export of prediction results and explanations in standard formats for research or clinical purposes.

3.2 NON-FUNCTIONAL REQUIREMENTS

The system shall maintain prediction accuracy above 75% for multi-class classification and process inputs within 5 seconds. It should handle concurrent analysis of multiple posts, ensuring stable performance under typical research and professional usage. Security measures, such as anonymization of sensitive text data, shall be implemented, alongside a user-friendly interface accessible across devices. System maintainability, logging, and version control shall also be supported to enable updates and monitoring of models and configurations.

3.3 SOFTWARE REQUIREMENTS

The development environment will use Python 3.8 or higher as the primary programming language. Machine learning models will be implemented using scikit-learn, while transformer-based architectures such as BERT and RoBERTa will be used through the Hugging Face Transformers library. Explainable AI outputs will be generated using SHAP to provide human-interpretable insights into model predictions for healthcare professionals and researchers.

The web application will be developed with React.js for the frontend and FastAPI for the backend API, which will handle model inference and predictions. Data persistence and session management during development and testing will use lightweight solutions such as SQLite or JSON-based storage. Jupyter Notebooks will be used for experimentation and analysis, and Git will manage version control.

3.4 DATASET REQUIREMENTS

The dataset shall comprise a minimum of 12,000 Reddit posts collected from 11 mental health-focused communities, covering all target categories. Each post will include a title, content, subreddit, category label, timestamp, and combined text field. Preprocessing will remove spam, duplicates, and personally identifiable information. Class weighting and validation splits shall ensure balanced representation across categories, supporting multi-class classification and reliable model evaluation.

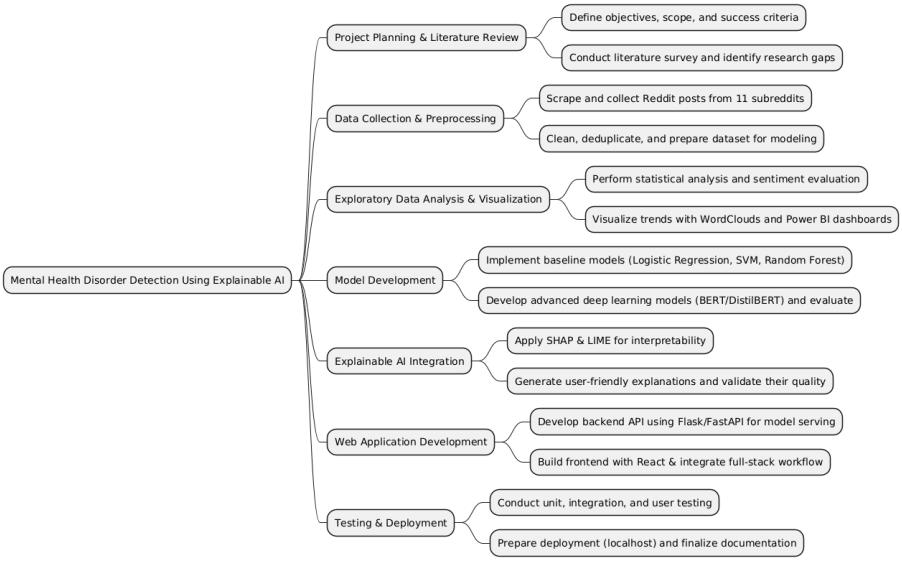
3.5 USER REQUIREMENTS / STAKEHOLDER NEEDS

Healthcare professionals shall receive interpretable explanations aligned with clinical understanding to support early detection and research. Researchers shall have access to batch processing, detailed performance metrics, and exportable results for analysis. The interface shall be accessible to non-technical users, providing clear explanations in plain language. Administrators shall have monitoring, logging, and update capabilities to maintain system performance. Ethical considerations, including anonymization and appropriate use disclaimers, shall guide all system interactions.

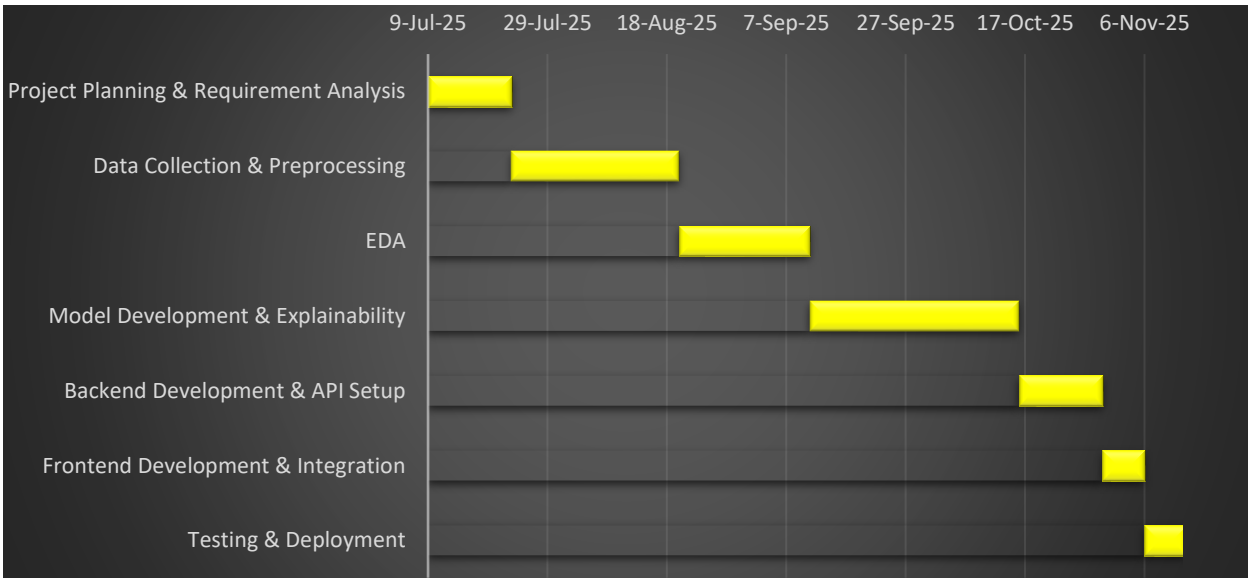
4. SYSTEM DESIGN AND PROJECT PLANNING

4.1 PROJECT PLANNING

The project planning phase outlines the structured approach for developing the mental health detection system. The Work Breakdown Structure (WBS) divides the project into major phases: data collection and preprocessing, exploratory data analysis, model development, explainable AI integration, web application development, testing and evaluation, and deployment. Each phase includes specific tasks such as Reddit data scraping, text cleaning, baseline and transformer model training, SHAP implementation, and user interface development.



A Gantt chart provides visual representation of the project schedule, showing start and end dates for each task with dependencies. Key milestones include dataset preparation completion, model training, explainable AI integration, and web application deployment. This timeline enables effective monitoring of overlapping tasks and resource allocation throughout the project.



Phase 1: Project Planning & Requirement Analysis

Timeline: 9th July 2025 – 20th July 2025

- Define the project scope and objectives.
- Gather requirements and finalize features.
- Map the system architecture and identify tools, frameworks, and datasets needed.
- Finalize deliverables and assign responsibilities.

Phase 2: Data Collection & Preprocessing

Timeline: 21st July 2025 – 18th August 2025

- Collect data from chosen
- Perform data cleaning: remove duplicates, handle missing values, normalize text.
- Merge data sources into a unified dataset.
- Prepare processed dataset for further analysis.

Phase 3: Exploratory Data Analysis (EDA)

Timeline: 19th August 2025 – 7th September 2025

- Perform statistical analysis and visualize data distributions.
- Identify correlations, patterns, and anomalies in the dataset.
- Extract useful features for machine learning models.
- Document insights for guiding model development.

Phase 4: Model Development & Explainability

Timeline: 8th September 2025 – 16th October 2025

- Develop machine learning models for disorder detection.
- Train and validate models using the cleaned dataset.
- Implement Explainable AI (SHAP or similar) to interpret predictions.
- Compare models and select the best-performing approach.

Phase 5: Backend Development & API Setup

Timeline: 17th October 2025 – 26th October 2025

- Develop backend logic to integrate model inference.
- Create REST APIs for model access.
- Ensure scalability and security for API endpoints.
- Connect to database for storage and retrieval of user/query data.

Phase 6: Frontend Development & Integration

Timeline: 27th October 2025 – 2nd November 2025

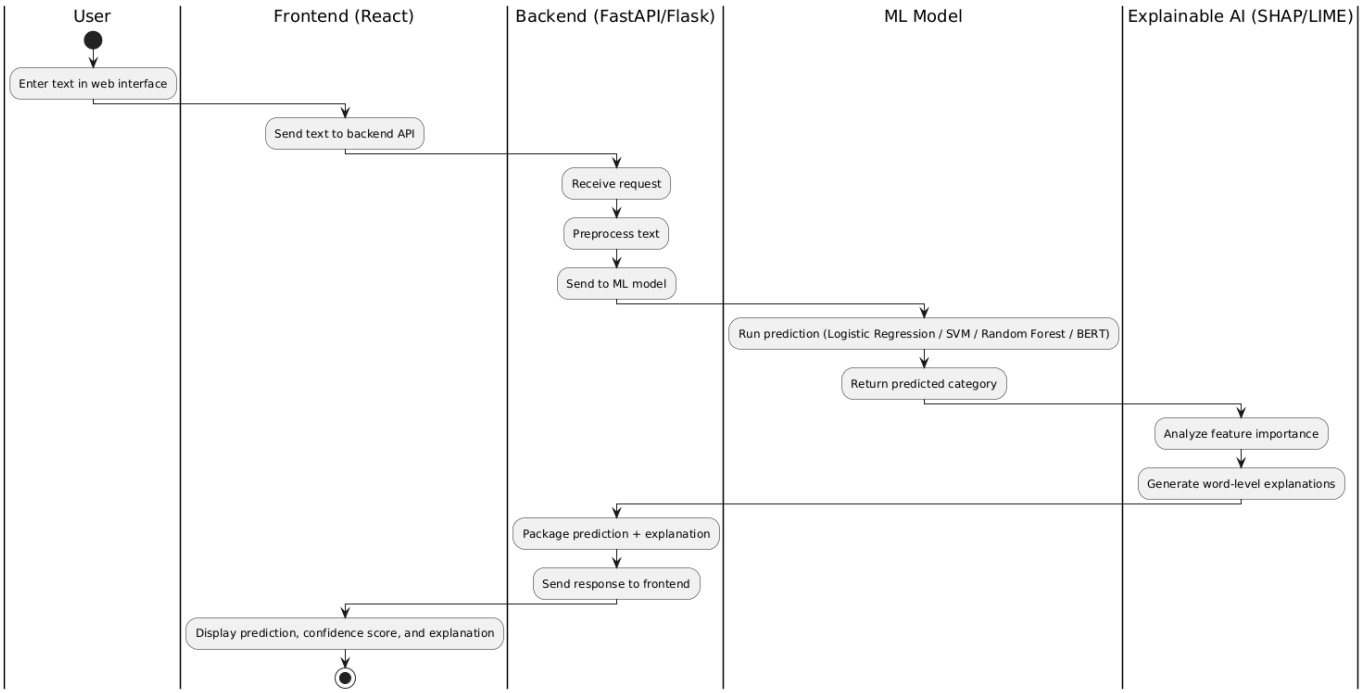
- Build the user interface for the application.
- Integrate frontend with backend APIs.
- Ensure seamless user experience with real-time model outputs.
- Add visual explanations for predictions (using SHAP plots, etc.).

Phase 7: Testing & Deployment

Timeline: 3rd November 2025 – 6th November 2025

- Perform unit, integration, and system testing.
- Conduct user acceptance testing (UAT).
- Optimize performance and fix bugs.
- Deploy final system for demonstration and usage.

An activity diagram illustrates the sequential and parallel flow of project activities, highlighting decision points such as model performance evaluation with feedback loops for retraining or fine-tuning as needed. Together, the WBS, Gantt chart, and activity diagram provide comprehensive planning that guides the project from initial data collection to final deployment, ensuring organized and efficient task execution.



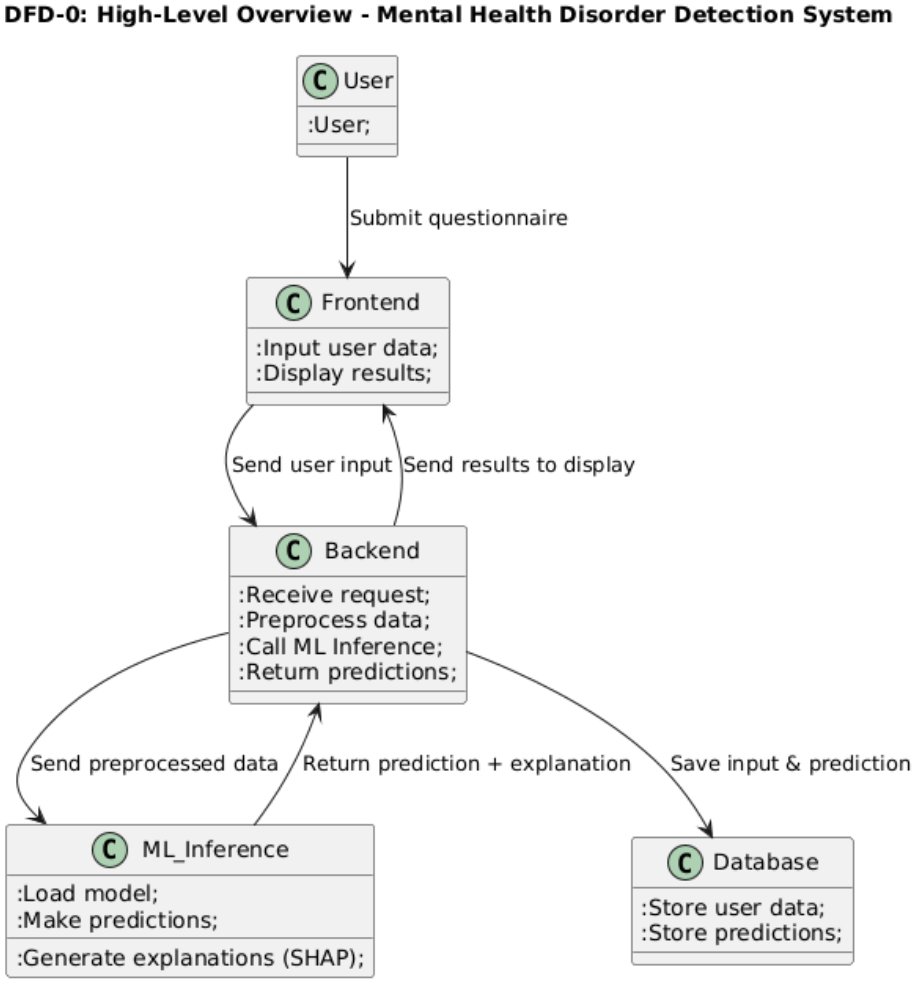
4.2 HIGH-LEVEL DESIGN

The High-Level Design provides an overview of the system architecture, focusing on the main components and their interactions rather than implementation details. The system is divided into three main layers: frontend, backend, and machine learning module. This modular separation ensures clarity, maintainability, and scalability, while keeping the system lightweight enough to run on standard hardware without requiring specialized infrastructure.

The frontend is developed with React and provides a user-friendly interface for submitting text, viewing predictions, and exploring explanations. Users see key outputs such as prediction confidence scores, highlighted important words, and visual summaries of feature importance. The interface is responsive and works across different devices and screen sizes, making the system accessible to both researchers and healthcare professionals.

The backend, implemented with FastAPI, acts as the central controller that coordinates data flow. It receives requests from the frontend, applies basic preprocessing to the text, forwards data to the models for inference, and returns predictions and explanations in a structured format. RESTful endpoints make integration straightforward, and the modular design allows future enhancements, such as adding more models or extending the number of mental health categories.

The machine learning module includes both traditional classifiers (Logistic Regression, SVM, Random Forest) and transformer-based models (BERT/DistilBERT) for multi-class classification of 11 mental health conditions. The explainability component integrates SHAP and attention visualizations, providing interpretable outputs that help users understand which words or phrases contributed most to each prediction. HLD emphasizes the overall structure and interaction of components rather than internal coding details.



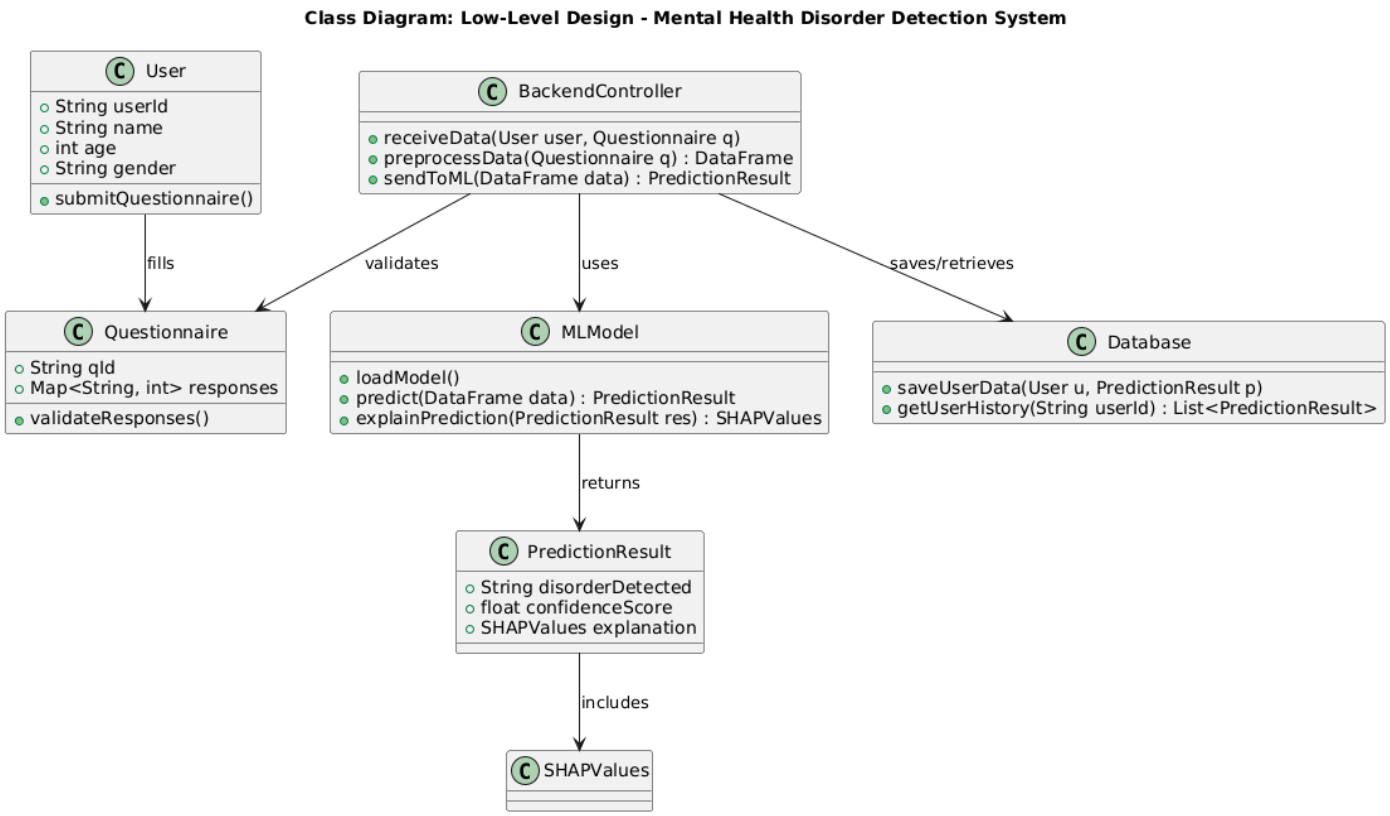
4.3 LOW-LEVEL DESIGN

The Low-Level Design focuses on the internal implementation of each component, detailing how the system actually functions. It describes modules, classes, methods, and data flows in a practical and achievable way.

The frontend module consists of components for text input, prediction display, and explanation visualization. Each component manages its state using React hooks, handles API errors, and renders results clearly for the user. Simple design patterns are used to ensure that the application remains lightweight and responsive without overcomplicating the interface.

The backend module is divided into functional services. A preprocessing service handles text cleaning and tokenization. The model service loads baseline and transformer models, performs predictions, and passes results to the explainability service. The explainability service computes SHAP values and attention maps and formats outputs into JSON for easy frontend consumption. Logging and error handling are included to maintain stability and allow debugging without heavy infrastructure.

The machine learning module includes classes for baseline models using scikit-learn and transformer-based models using Hugging Face. A unified interface standardizes training, inference, and explanation extraction. Preprocessing pipelines, model loading, and inference are implemented as modular functions for easy updates.



5. REFERENCES

1. MDHAN (Zogan et al., 2020)

Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2020). Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. *arXiv preprint arXiv:2007.02847*.

2. Ensemble transformer-based approaches (2024)

Zhang, et al. (2024). Psychiatric disorders from EEG signals through deep learning models. *Biomedical Signal Processing and Control*.

3. Social media mental health studies (2023-2024)

Garg, M., et al. (2023). Mental health analysis in social media posts: A survey. *PMC Articles*, PMC9810253.

4. Psychiatric disorder detection models (2024)

Author (2024). Multichannel convolutional transformer for detecting mental disorders. *Nature Scientific Reports*, Article s41598-025-98264-w.

5. Systematic XAI evaluations (Kerz et al., 2023)

Kerz, E., Zanwar, S., Qiao, Y., & Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry*, 14, Article 1219479