INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

Machine Learning-Based Mental Health Classification System: Design, Implementation, and Evaluation

S.A. Adefemi[1], A.D. Ojo[2], A.A. Alabi[3], A. Tewogbade[4], O. Ikotun[5]

[1,3,4,5]Department of Computer Science, Osun State University, Osogbo.

[2]Department of Computer Science, Osun State Polytechnic, Iree.

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.11030034

Received: 07 March 2026; Accepted: 16 March 2026; Published: 02 April 2026

ABSTRACT

Mental disorders such as Bipolar Type-1, Bipolar Type-2, and Depression continue to affect millions of people worldwide, yet early and accurate diaganosis is challenging due to stigma, limited resources, and the subjectivity of self-reporting. Trying to bridge this gap, this project sought to develop a mental health diagnosis system that possesses the ability to classify individuals into Bipolar Type-1, Bipolar Type-2, Depression, or Normal states based on organized user input. Utilizing data acquired from an online repository, the system was designed with careful data cleaning, pre-processing, and class balancing using SMOTE for equal representation. Logistic Regression, Decision Tree, and Random Forest models were trained individually and then ensemble together using both hard and soft voting ensemble methods to obtain more stable predictions. The final ensemble model outperformed the individual models with accuracy up to 80%. This solution was deployed as a simple web app where users are able to answer a few guided questions and receive AI-generated feedback about their possible mental state instantly. The project demonstrates that the application of an ensemble of machine learning models will enhance early mental health screening and provide a supportive, accessible tool that will encourage individuals to seek professional help when needed.

Keywords: Bipolar Disorder, Depression, Diagnosis System, Ensemble Learning, Machine Learning, Mental Health.

INTRODUCTION

Mental disorders, including depression and bipolar illnesses, are rapidly emerging as critical global public health concerns. Studies indicate that more than one in eight individuals worldwide experience mental health challenges, yet many do not receive adequate care or diagnosis due to stigma, limited awareness, or restricted access to mental health professionals (Pim, Afzal, & Kamaldeep, 2023). Early diagnosis and intervention are essential for improving the quality of life of affected individuals and reducing the long-term social and economic burden of these disorders. With the expansion of digital technologies, artificial intelligence (AI) and machine learning (ML) have demonstrated significant potential in addressing complex healthcare challenges (Arrieta et al., 2020). In particular, Natural Language Processing (NLP) enables computers to process and understand human language, allowing for the extraction of emotional and psychological insights from text sources such as social media posts, journal entries, or structured self-reports (Zaagsma, Stökl Ben Ezra, Rürup, Margolis, & Levi, 2022). Leveraging NLP in mental health research facilitates the detection of early signs of psychological disorders through analysis of user-generated text data, providing a scalable and non-invasive screening approach.

Despite these advancements, significant challenges remain. A substantial proportion of individuals with depression or bipolar disorders remain undiagnosed or misdiagnosed due to social stigma, inaccessibility of qualified mental health professionals, and the inherent subjectivity of conventional diagnostic methods (Osborn et al., 2020; Alzheimer's Association, 2023). Furthermore, many existing AI-based diagnostic systems suffer from limitations such as small or biased datasets, model overfitting, poor generalizability, and reliance on single-model approaches that reduce predictive accuracy and reliability (Bickman, 2020). These gaps highlight the need for a robust, accurate, and deployable system capable of performing first-level mental health screening using structured user input. This study proposes the development of an ensemble-based machine learning system designed to detect mental health conditions specifically bipolar I, bipolar II, depression, and normal states by analyzing structured user input. The system integrates multiple classification models, including Logistic Regression, Random Forest, and Decision Tree, combined through a Voting Ensemble method to enhance accuracy and minimize vulnerabilities such as overfitting and bias. The final model was deployed as a web-based application, allowing users to enter symptom-related data and receive immediate diagnostic feedback. This approach aimed to provide an accessible, cost-effective, and non-invasive tool that can serve as an ancillary resource for early intervention, self-assessment, and mental health awareness, particularly in regions with limited access to professional care.

The objective of this project was to develop a machine learning-based system for primary diagnosis of mental health conditions, specifically bipolar 1, bipolar 2, depression, and normal mental status based on user questionnaire selections. Three machine learning algorithms will be implemented in this project: Logistic Regression, Random Forest and Decision Tree. The models were combined using a Voting Ensemble method to improve classification correctness and reliability. The system was implemented in Python using standard machine learning and natural language processing libraries and deployed through a web-based interface for usability.

LITERATURE REVIEW

Mental disorders such as depression and bipolar illnesses remain pressing global public health concerns, with more than one in eight individuals worldwide experiencing mental health challenges. Despite the prevalence, many individuals remain undiagnosed or misdiagnosed due to stigma, limited awareness, and inadequate access to trained mental health professionals (Pim, Afzal, & Kamaldeep, 2023). Early recognition and intervention have been shown to improve the quality of life of affected individuals and reduce the long-term social and economic burden of untreated conditions. However, traditional diagnostic practices rely heavily on clinician-administered assessments and self-reported symptoms, which are inherently subjective and often fail to provide timely detection, especially in underserved regions (Osborn et al., 2020; Alzheimer's Association, 2023). This diagnostic gap underscores the need for innovative, scalable, and objective approaches to mental health detection. The growing influence of digital technologies and the emergence of Artificial Intelligence (AI) and Machine Learning (ML) have opened new opportunities in healthcare applications, including mental health diagnosis. In particular, Natural Language Processing (NLP) allows computers to interpret and analyze human language, extracting emotional and psychological cues from diverse text sources such as social media posts, journal entries, and structured self-reports (Zaagsma, Stökl Ben Ezra, Rürup, Margolis, & Levi, 2022). NLP-based approaches have demonstrated potential as non-invasive and scalable tools for early detection of mental health conditions (Arrieta et al., 2020). A comprehensive survey of NLP in mental health diagnosis highlighted its growing adoption but also noted challenges such as language dependence, small sample sizes, and limited deployment in real-world clinical settings (Computer Science Review, 2024). Ensemble learning techniques have also been increasingly employed to improve the robustness of AI-based mental health detection systems. Studies have shown that combining classifiers such as Random Forest, AdaBoost, XGBoost, and Voting Classifiers enhances prediction accuracy compared to single-model approaches. For instance, Aslan (2024) demonstrated that Random Forest achieved the highest predictive performance on a dataset of over 124,000 tweets, with the Voting Classifier performing as a strong secondary option. Bokolo and Liu (2023) further compared conventional machine learning methods with transformer-based models such as RoBERTa, achieving accuracies as high as 98% in depression detection from Twitter data. These findings highlight that ensemble methods can effectively mitigate overfitting and leverage the strengths of multiple models.

Recent research has also expanded into transformer architectures and domain-specific large language models (LLMs) for mental health analysis. Cheng et al. (2024) developed DORIS, a system combining LLMs with medical knowledge-guided features to detect depression from social media, achieving both high accuracy and interpretability. Similarly, mental health-specific transformer models such as MentalBERT, MentalRoBERTa, PHS-BERT, DisorBERT, and MentaLLaMA have demonstrated superior performance in identifying depressive expressions compared to general-purpose LLMs (arXiv, 2024). Multimodal approaches that combine text with visual or audio features have also shown promise. Bucur et al. (2023) proposed a time-enriched multimodal transformer leveraging text and image embeddings, achieving F1-scores above 0.93, while Ghadiri et al. (2022) explored the integration of voice signal features with textual embeddings to enhance depression detection. Cross-cultural generalization remains a critical challenge for AI-driven mental health tools. Nuredin Ali et al. (2024) demonstrated that models trained predominantly on data from the Global North often fail to generalize effectively to populations in the Global South, exposing biases that can reduce detection accuracy. Likewise, research on Arabic NLP for mental health revealed that specialized models such as AraBERT achieved over 90% accuracy in detecting depression and suicidal ideation, significantly outperforming conventional classifiers (MDPI, 2024). These studies emphasize the need for diverse and balanced datasets to ensure equitable performance across languages and cultures.

Additionally, literature shows that AI and NLP have revolutionized the potential for early mental health screening, with ensemble models and transformer-based approaches delivering strong performance. Nonetheless, challenges persist in generalization, bias mitigation, and explainability, particularly in cross-cultural contexts. Current systems are often limited by single-model approaches, small or biased datasets, and a lack of real-time, user-facing deployment. These gaps justify the development of an ensemble-based machine learning system that integrates models such as Logistic Regression, Random Forest, and Decision Tree using a Voting Ensemble. When deployed as a web-based platform for structured user input, such a system has the potential to deliver accurate, robust, and interpretable first-level screening for depression and bipolar disorders, complementing traditional clinical evaluations and increasing access to early mental health support.

METHODOLOGY

This study followed a structured methodology to design, develop, and deploy a web-based mental health diagnosis system powered by machine learning. The methodology encompassed dataset collection and preprocessing, classification algorithm selection, model evaluation, system development, and workflow modeling, ensuring the resulting system was accurate, reliable, and user-friendly.

Dataset Collection and Preprocessing

The system utilized a publicly available dataset from Kaggle containing 121 instances and 18 features representing mental health questionnaire responses. The features included both categorical and numerical data, with three numeric features: Sexual Activity, Concentration, and Optimism. Preprocessing steps were essential to prepare the data for machine learning. Categorical variables were label-encoded to convert textual data into numerical form suitable for model training. Additionally, Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance within the dataset (Sarker, 2021). SMOTE generates synthetic samples for underrepresented classes, such as Bipolar Type-1, Bipolar Type-2, and Depression, thereby ensuring the training data is balanced.

Classification Algorithms

The methodology also involved the selection and implementation of suitable classification algorithms to predict mental health conditions accurately. Three base classifiers were selected based on their performance and interpretability: Logistic Regression, Random Forest, and Decision Tree. Logistic Regression was chosen for its efficiency in high-dimensional feature spaces and strong performance in classification tasks. Random Forest, a tree-based ensemble method, was selected for its robustness, ability to handle mixed data types, and resistance to overfitting. Decision Tree was implemented for its non-parametric nature and high interpretability, a crucial feature in sensitive domains such as mental health where model outputs must be explainable. To enhance overall performance and minimize the weaknesses of individual models, a Voting Ensemble was adopted. This ensemble combined the predictions of the three base models using hard voting (majority class decision) and soft voting (probability averaging), resulting in improved accuracy, stability, and reliability.

Model Evaluation

To ensure the predictive reliability and usability of the system, the trained models were evaluated using widely accepted performance metrics: accuracy, precision, recall, and F1-score. These metrics provided a comprehensive understanding of each model's predictive correctness and sensitivity to different classes. Evaluation was particularly important given the multi-class classification problem, which involved classifying users into Bipolar Type-1, Bipolar Type-2, Depression, or Normal. Confusion matrices were also generated to provide a detailed breakdown of model predictions for each class, enabling the identification of misclassifications and demonstrating how the Voting Ensemble improved upon the individual classifiers.

System Development

The mental health diagnosis system was implemented as a modular web-based application consisting of a frontend for user interaction and a backend for data processing and model inference. The frontend was developed using HTML, CSS, and JavaScript to ensure clarity, responsiveness, and ease of use. It hosted a structured questionnaire containing 17 required inputs, capturing emotional, psychological, and behavioral indicators of mental health. Upon completion, inputs were validated and transmitted to the backend via HTTP requests, and the diagnostic results were presented in a concise, non-alarming format for easy interpretation. The backend was implemented in Python using the Flask framework to manage server-side operations and machine learning integration. Its responsibilities included receiving user input, preprocessing data to align with the trained model's expected format, loading the pre-trained models stored with Pickle, and generating predictions using the Voting Ensemble. The final classification Bipolar Type-1, Bipolar Type-2, Depression, or Normal was sent back to the frontend as JSON data and stored in a MySQL database for record-keeping.

System Architecture and Workflow

The system architecture was designed to clearly separate client-side and server-side operations, ensuring smooth user interaction and real-time processing. The frontend handles data collection, while the backend processes inputs and performs classification. To visualize functional interactions, a use case diagram was created. The primary actor is the general user, while an optional admin actor is reserved for future functionalities such as model updates and data monitoring. The core use cases include completing the questionnaire, submitting responses, and receiving predictions as shown in figure 1.

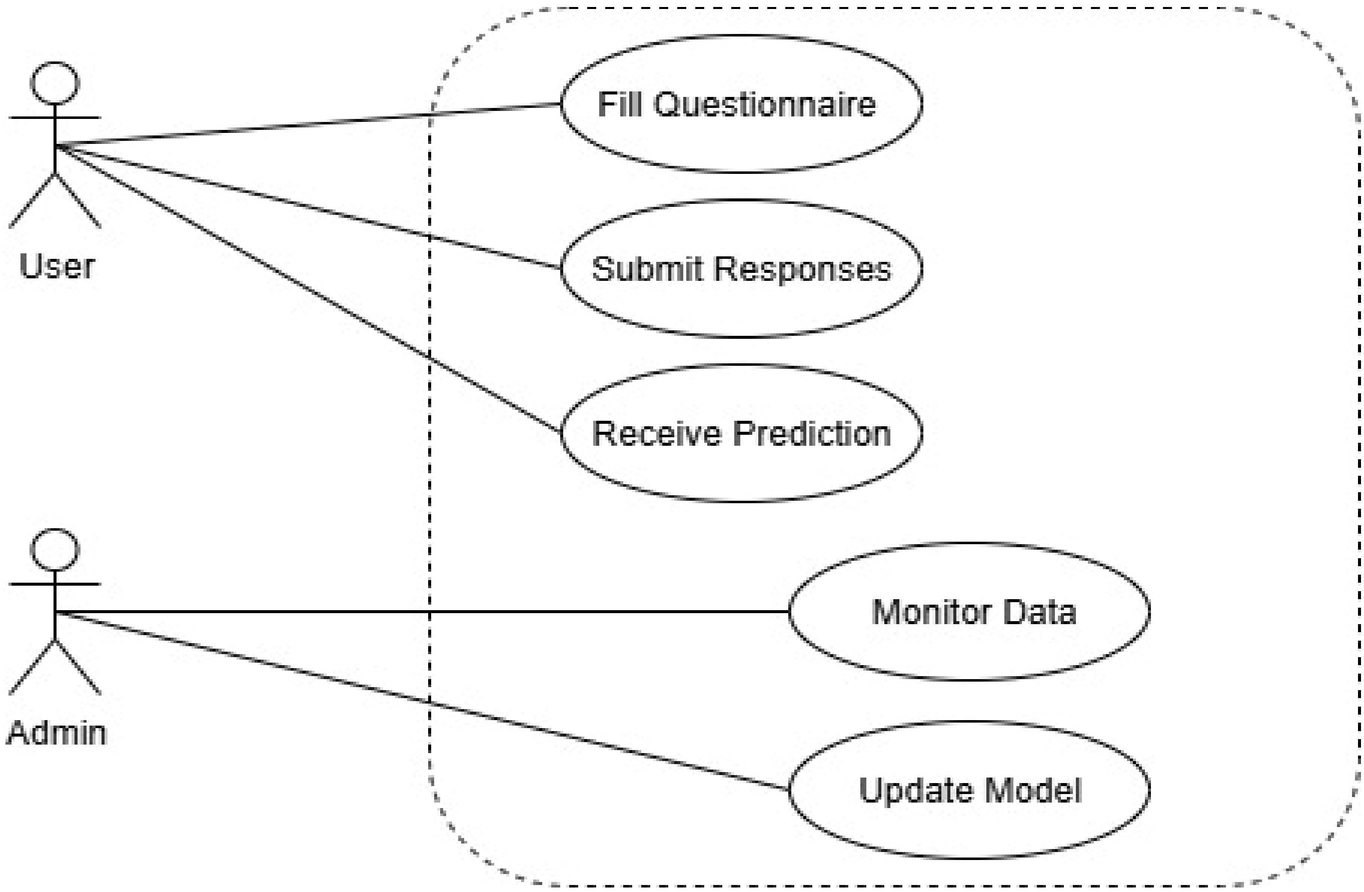

Figure 1: Use Case Diagram

Discussion

The dataset was first checked for missing values, and none were found. Categorical features and the target variable (Expert Diagnose) were label-encoded to convert them into numerical values suitable for machine learning algorithms. No feature scaling was applied, as the features were mostly categorical. To address class imbalance, SMOTE was used to oversample the minority classes in the training set. The pre-processed data was then split into training and testing sets using stratified sampling to preserve class distribution.

Feature Importance Analysis

To better understand which features contributed most to the model's predictions, feature importance was analyzed using the Random Forest model. Figure 2 illustrates the feature importance derived from a trained model, likely a tree-based classifier such as a Random Forest or Decision Tree. These importance scores indicate how much each feature contributes to the predictive power of the model. From the chart, "Mood Swing" emerges as the most influential feature, with an importance score of approximately 0.24, suggesting it plays a crucial role in the classification task. This is followed by "Optimism" and "Sexual Activity", both of which also show significant influence. Other notable features include "Euphoric", "Suicidal Thoughts", and "Concentration", each contributing moderately to the model's decision-making process. On the lower end of the importance spectrum are features such as "Admit Mistakes", "Overthinking", and "Try-Explanation", indicating that they have minimal impact on the model's output.

Model Training

Three major classification models were developed in this study: Logistic Regression, Random Forest, and Decision Tree. To enhance the performance of the models, hyperparameter tuning was performed using GridSearchCV to enable the choice of the most appropriate combination of parameters for each of the classifiers. To resolve the issue of class imbalance in the data set, the training set was balanced using SMOTE (Synthetic Minority Over-sampling Technique) prior to training. The performance and generalization of each individual model were evaluated on an independent test set. In addition to the evaluation of individual models, two ensemble techniques hard voting and soft voting were employed to combine the strengths of the classifiers.

Model Results

The results of this study provide a comparative assessment of five classification models Logistic Regression, Random Forest, Decision Tree, Soft Voting Ensemble, and Hard Voting Ensemble for the task of mental health condition classification. All models were trained on a pre-processed and SMOTE-balanced dataset and evaluated using accuracy, precision, recall, and F1-score. Confusion matrices were used to analyze class-specific prediction performance and identify misclassification patterns.

The Logistic Regression model achieved an overall accuracy of 69%, with a precision of approximately 70% and recall of 69%. While it provided reasonable predictive performance, the model showed a bias toward majority classes, as reflected in its confusion matrix (Figure 3), where minority classes were less accurately predicted. The Random Forest classifier outperformed the individual models with an accuracy of 79%, and both precision and recall exceeding 79%. Its confusion matrix (Figure 4) demonstrated strong classification performance across both majority and minority classes, benefiting from its ensemble of decision trees that reduce bias and enhance generalization. The Decision Tree model achieved 77% accuracy, with precision and recall of 78% and 77%, respectively. As shown in its confusion matrix (Figure 5), it performed fairly well but misclassified more instances in minority classes compared to Random Forest. Its single-tree structure limited its ability to generalize compared to the ensemble approach. The Soft Voting Ensemble yielded the best overall performance, achieving 80% accuracy, 82% precision, 80% recall, and an F1-score of 79%. By aggregating probability estimates from the individual models, it produced more balanced predictions across all classes. The confusion matrix (Figure 6) shows that this model improved classification for both majority and minority classes, mitigating the bias observed in single classifiers. Finally, the Hard Voting Ensemble also improved upon the base models, with an accuracy of 78%, precision of 80%, recall of 78%, and an F1-score of 77%. Its confusion matrix (Figure 7) indicates performance patterns similar to Random Forest and Decision Tree, effectively classifying

INTERNATIONAL JOURNAL OF RESEARCH AND INNOVATION IN APPLIED SCIENCE (IJRIAS)
ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

most majority class instances but slightly less balanced than the Soft Voting Ensemble due to its reliance on majority voting rather than probability weighting.

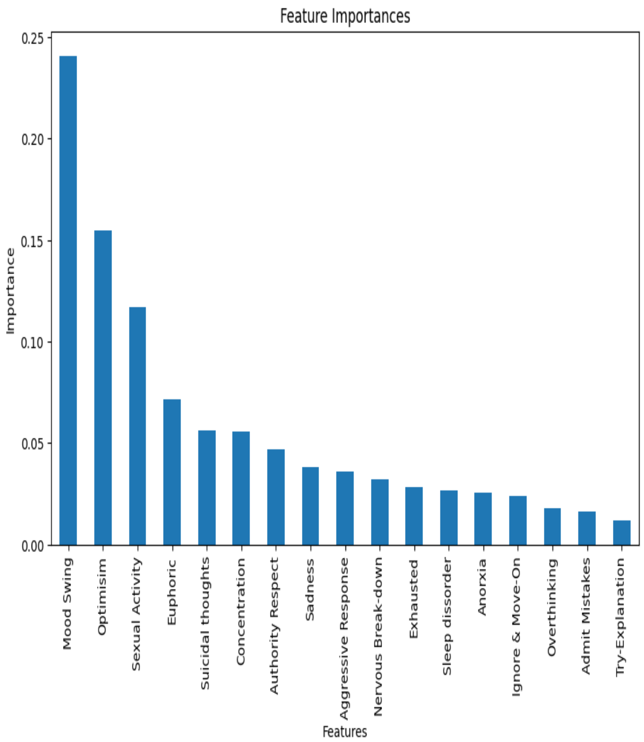

Figure 2: Feature Importance Determined by Random Forest Model.

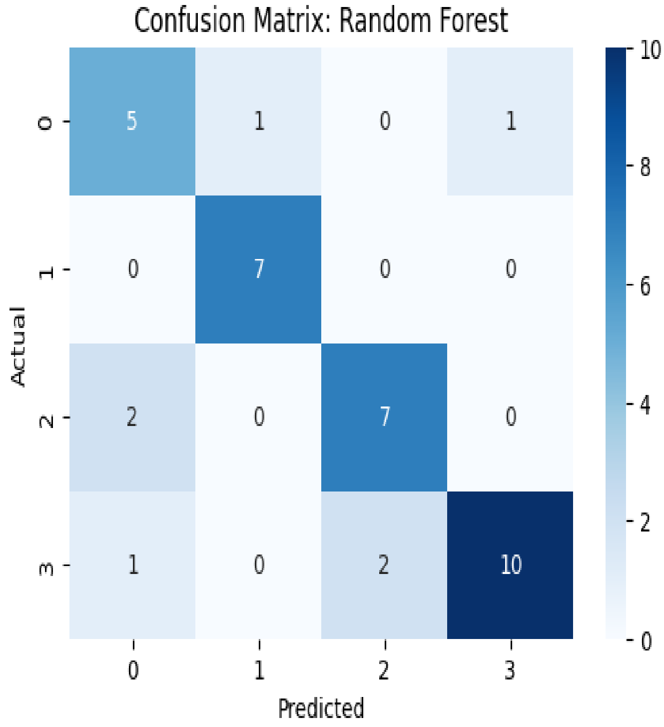

Figure 4: Confusion matrix for Random Forest

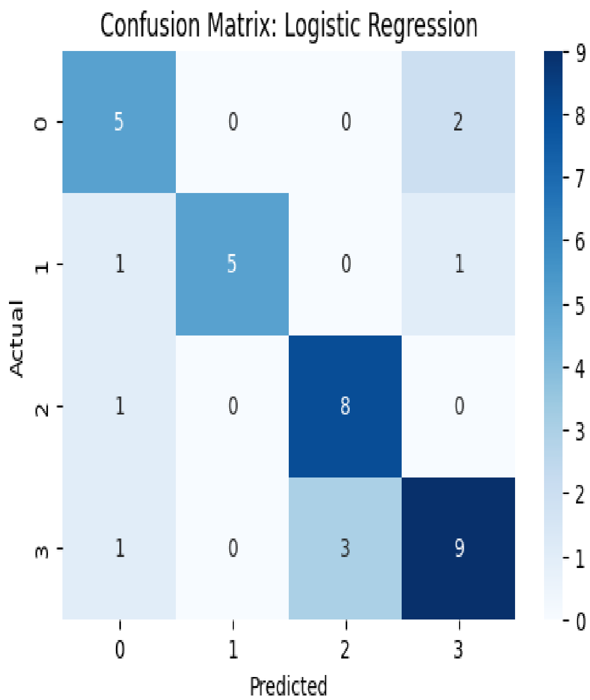

Figure 3: Confusion matrix for Logistic Regression

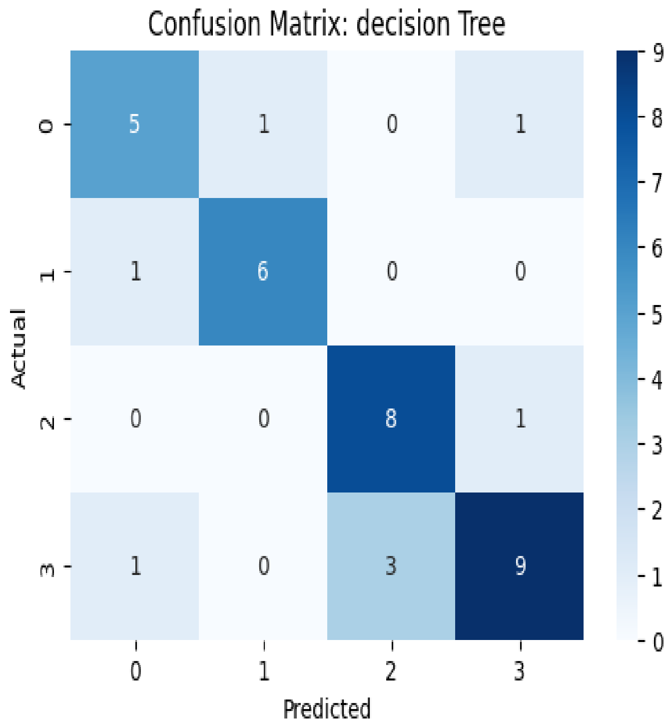

Figure 5: Confusion matrix for Decision Tree

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

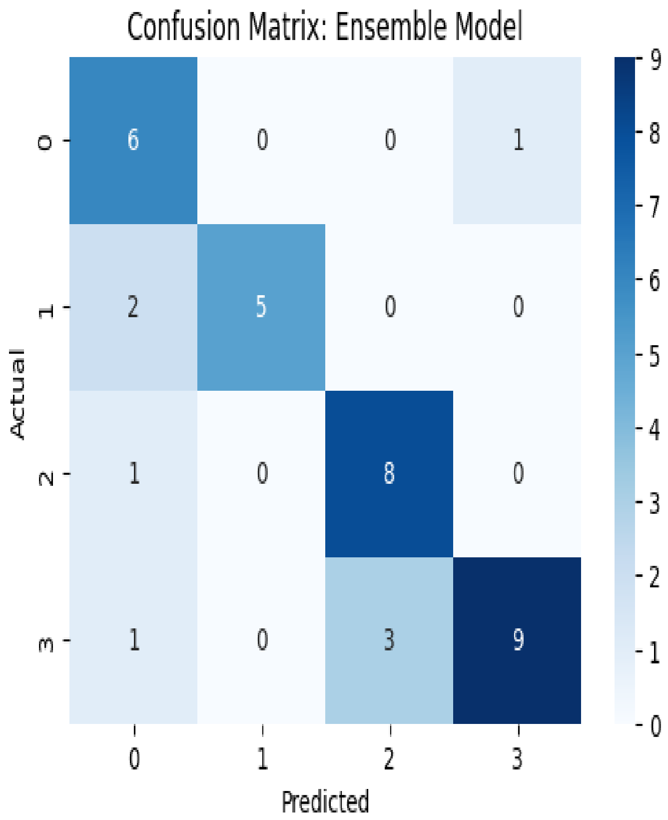

Figure 6: Confusion matrix for Soft Voting Ensemble

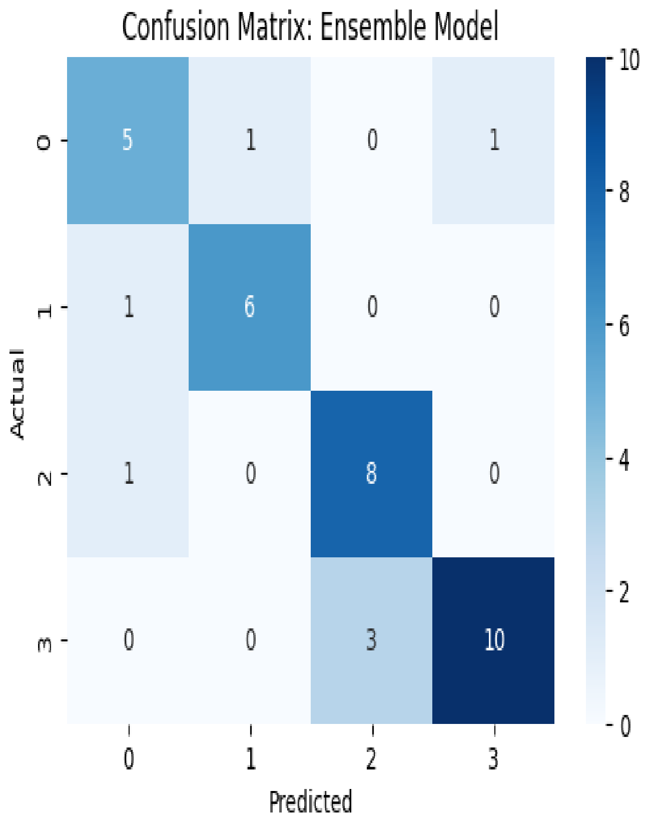

Figure 7: Confusion matrix for Hard Voting Ensemble

Model Comparison

A comprehensive performance comparison was carried out between the individual classifiers Logistic Regression, Random Forest, and Decision Tree and the ensemble approaches, namely Soft Voting and Hard Voting. While the individual models demonstrated varying degrees of accuracy and class-specific performance, the ensemble methods were designed to leverage the collective strengths of these base learners. Specifically, Soft Voting improves predictive performance by averaging the predicted probabilities of each model, whereas Hard Voting makes predictions based on the majority class votes among the classifiers. The results, summarized in Table 1, present the evaluation of all models on the test set using key metrics: Accuracy, Precision, Recall, and F1-Score, providing a clear insight into the effectiveness and robustness of each modelling approach.

Table 1: Performance Metrics for Individual and Ensemble Models on the Test Set

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.75	0.76	0.75	0.75
Random Forest	0.80	0.81	0.79	0.80
Decision Tree	0.77	0.78	0.77	0.76
Soft Voting Ensemble	0.80	0.82	0.80	0.79
Hard Voting Ensemble	0.78	0.80	0.78	0.77

The bar chart (Figure 8) visually highlights the superior performance of the ensemble and tree-based models, making it easy to identify the most effective approaches for this classification task.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

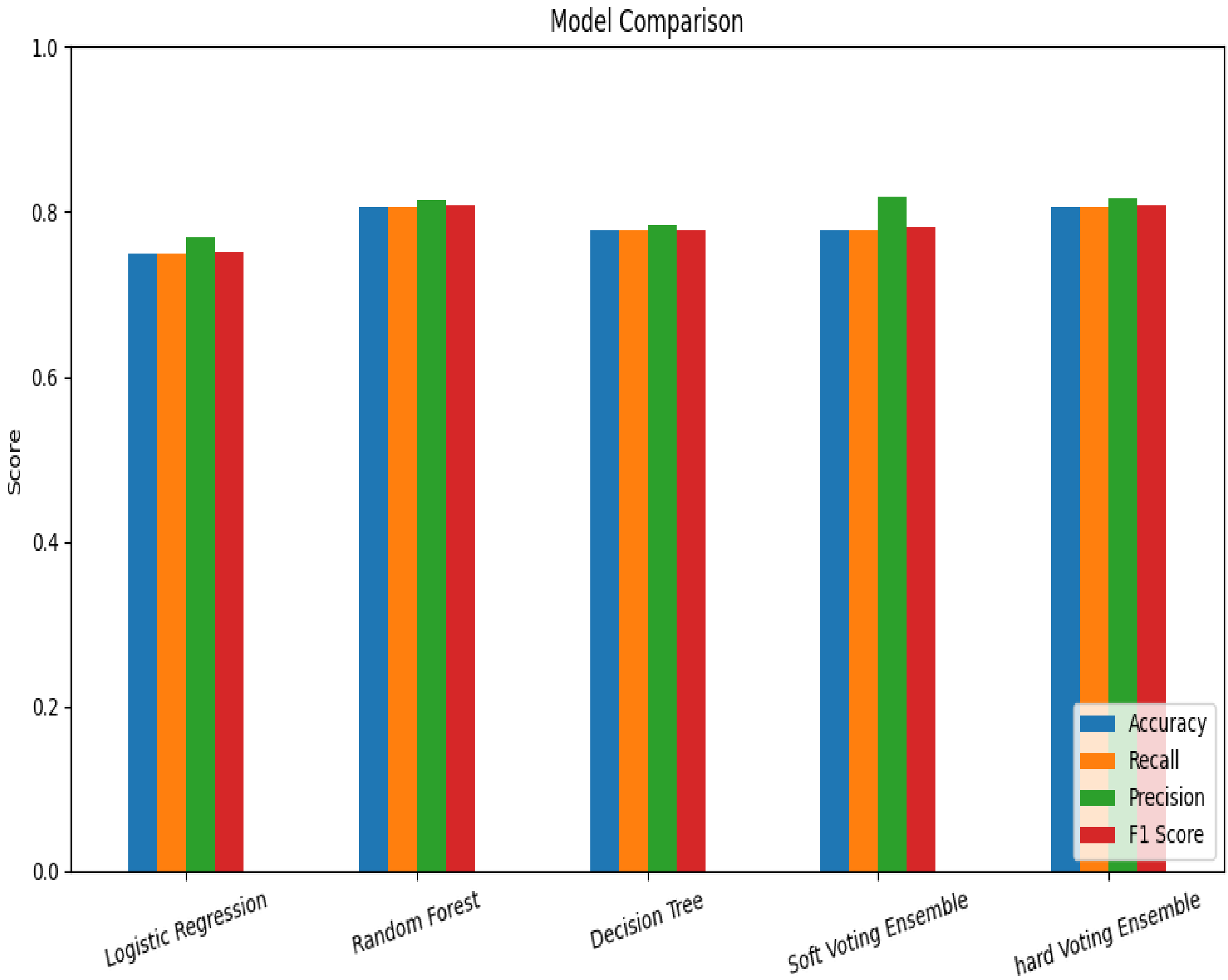

Figure 8: Model Comparison.

System Implementation and Sample Results

The final system was implemented as a web application using Flask, allowing users to input survey responses and receive a predicted mental health condition. The user interface features dropdown menus for all categorical features, ensuring valid input and a smooth user experience. The backend processes the input, maps it to the correct label-encoded values, and uses the trained ensemble model to generate a prediction, which is then mapped back to the original class name for clarity. Figures 9 to 14 illustrate the implementation of the mental health diagnosis system. Figures 9 and 10 display the user input interface, where users complete the structured multi-page questionnaire capturing emotional, psychological, and behavioral indicators. Figures 11 to 14 present sample output screens, demonstrating how the diagnostic results are displayed to the user in a clear, non-alarming, and easy-to-understand format. These outputs provide immediate feedback, indicating whether the user is classified as Bipolar Type-1, Bipolar Type-2, Depression, or Normal, thereby reflecting the seamless integration of the frontend interface with the machine learning backend.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

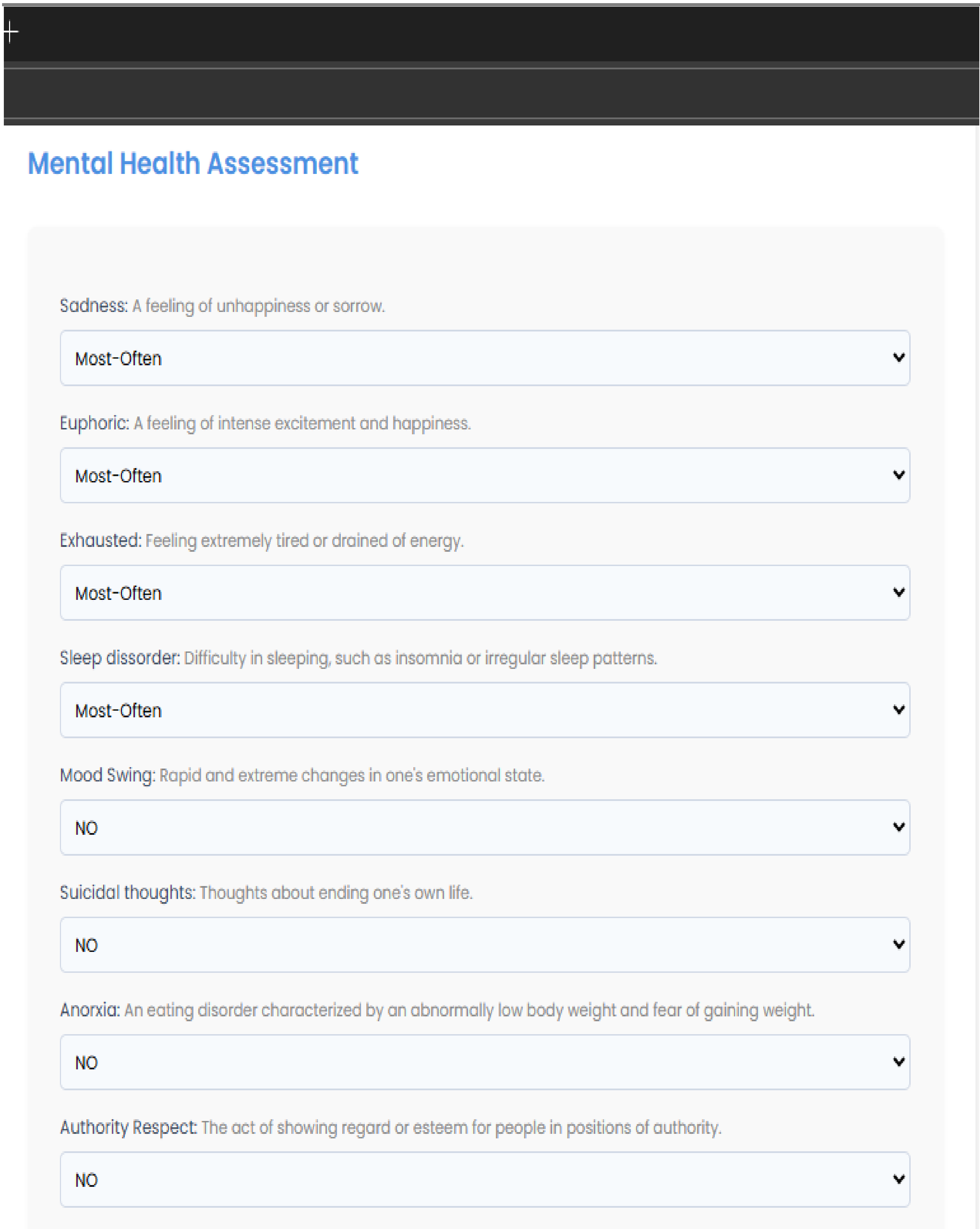

Figure 9: First page of the mental health diagnosis system input form.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

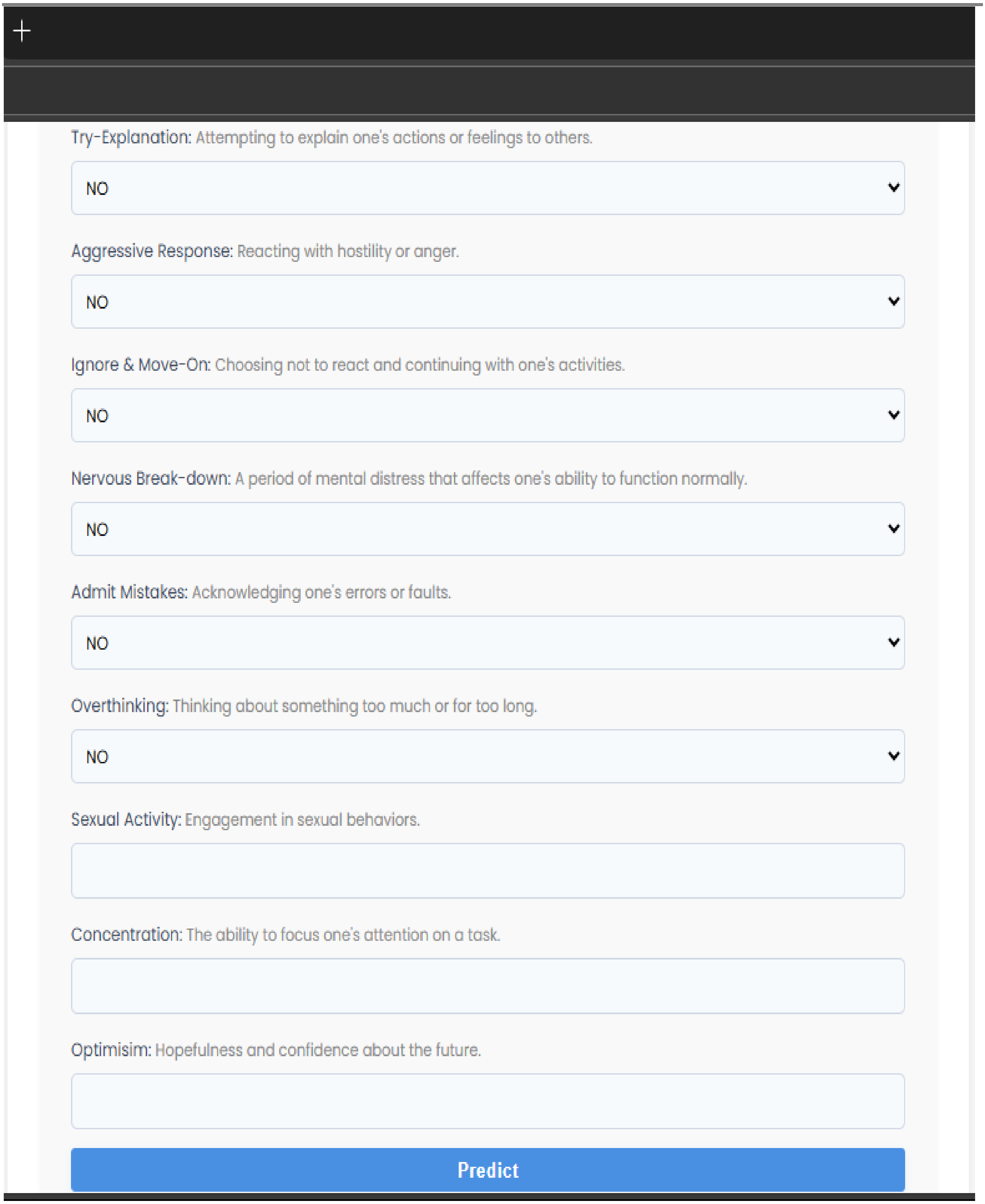

Figure 10: Second page of the mental health diagnosis system input form

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

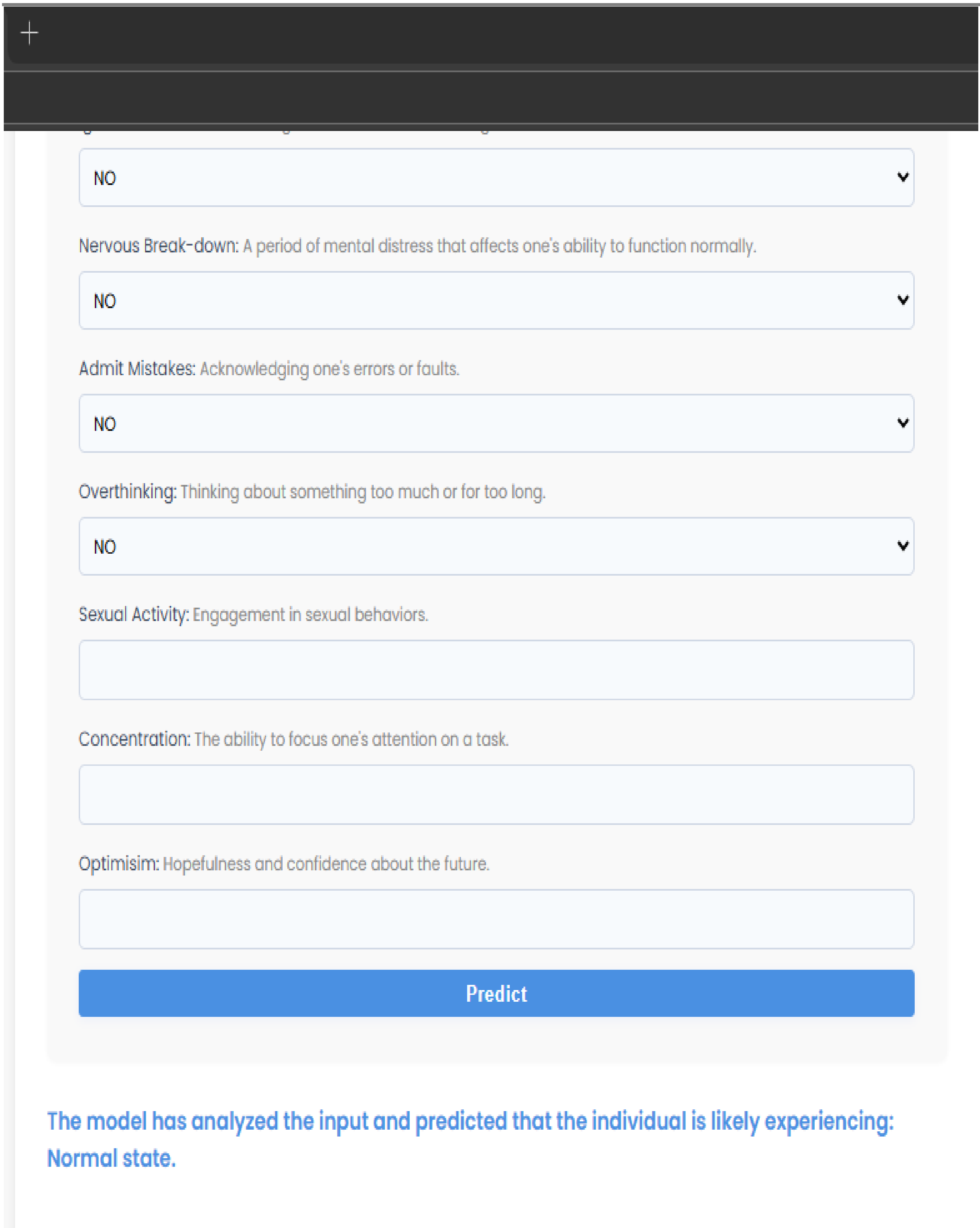

Figure 11: Sample diagnostic result displayed for a user input.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

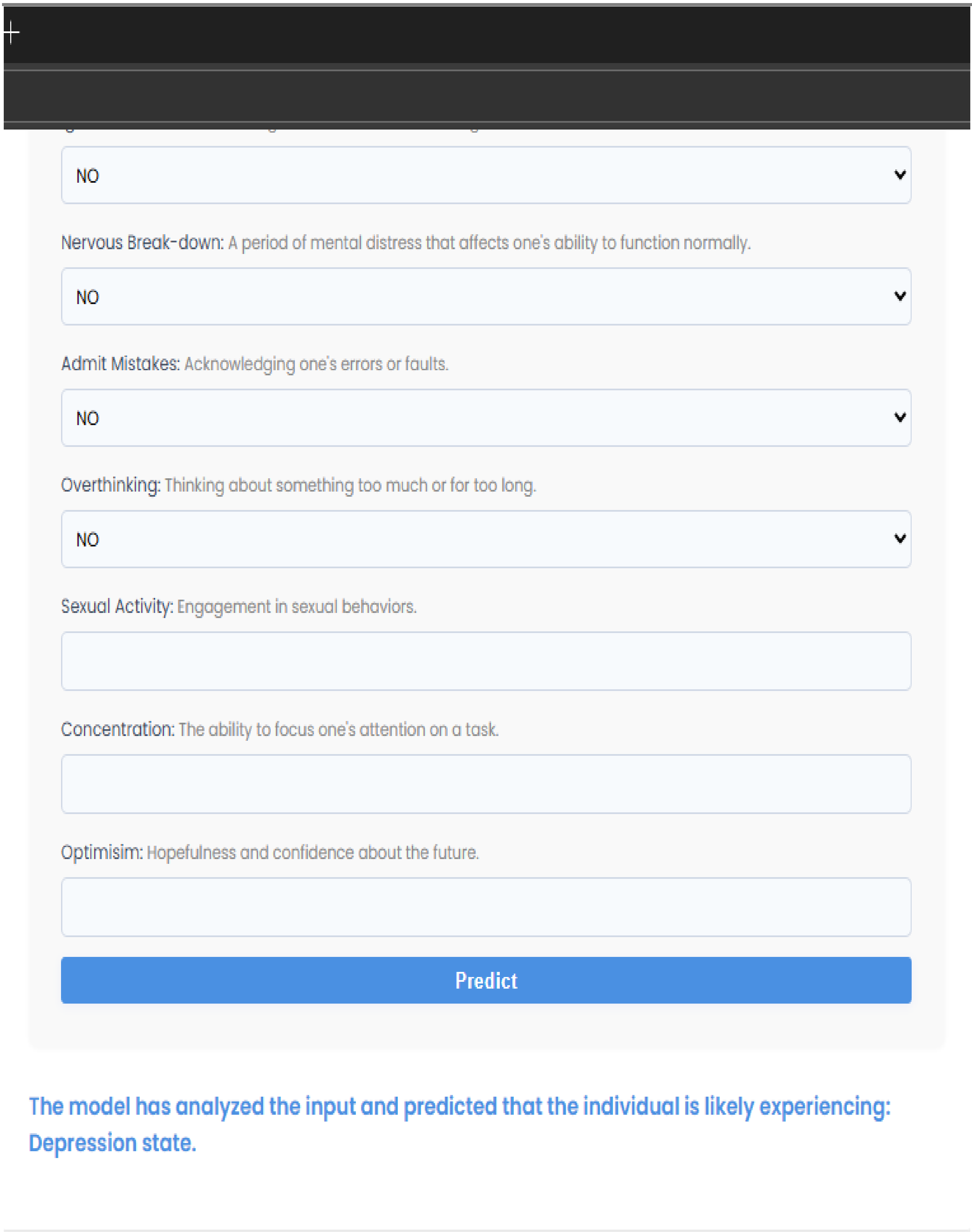

Figure 12: Example of system output showing predicted mental health status.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

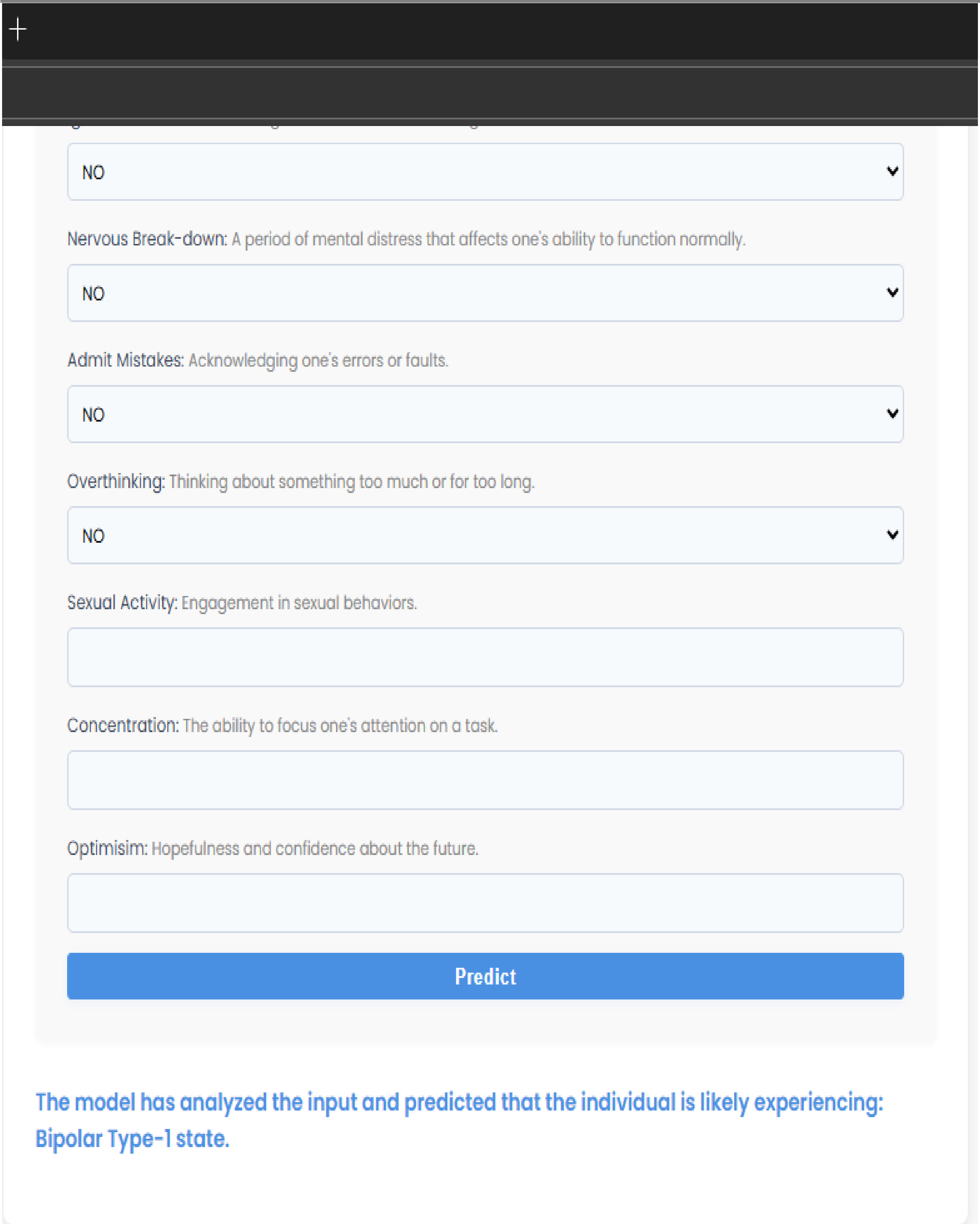

Figure 13: Diagnostic result page highlighting classified mental health category.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue III March 2026

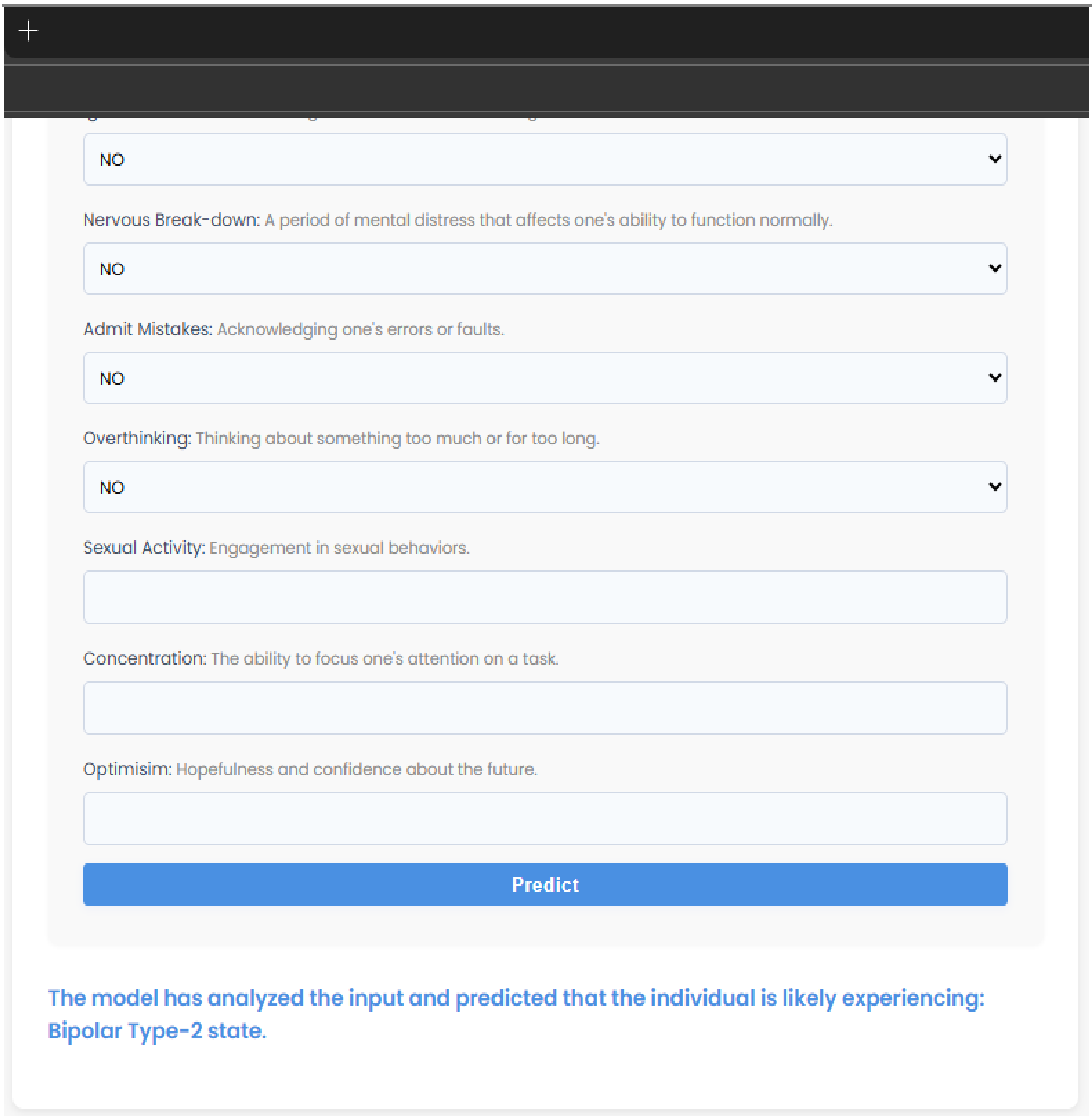

Figure 14: Final result display screen of the mental health diagnosis system.

CONCLUSION

This study successfully developed and evaluated a machine learning-based mental health diagnosis system capable of classifying users into Bipolar Type-1, Bipolar Type-2, Depression, or Normal categories based on structured self-reported inputs. By leveraging data preprocessing with SMOTE, ensemble learning techniques, and a web-based deployment framework, the system provides a scalable, accessible, and non-invasive first-level mental health screening tool.

The experimental results demonstrated that while individual models such as Logistic Regression, Random Forest, and Decision Tree achieved reasonable classification accuracy, their performances varied across classes, with minority classes being more prone to misclassification. The integration of these models through ensemble learning significantly enhanced overall performance, with the Soft Voting Ensemble achieving the highest accuracy (80%), along with balanced precision, recall, and F1-scores. Confusion matrix analyses further confirmed the ensemble model's ability to mitigate class bias and improve generalization across all mental health conditions. The deployment of the final model within a modular web application allows real-time user interaction, immediate diagnostic feedback, and a pathway for broader mental health awareness. Although this system is not a replacement for clinical evaluation, it can serve as an ancillary tool for early detection, self-assessment, and mental health education, especially in regions with limited access to professional mental health services.

Future work could focus on expanding the dataset to improve model generalization, incorporating natural language inputs for more flexible assessments, and enhancing interpretability to support clinical decision-making. By combining robust machine learning techniques with user-friendly web deployment, this study demonstrates the potential of AI-driven systems to support early mental health interventions and reduce the societal burden of psychological disorders.

REFERENCES

1. Alzheimer's Association. (2023). 2023 Alzheimer's disease facts and figures. Alzheimer's & Dementia, 19(4), 1598-1695. https://doi.org/10.1002/alz.13016
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012
3. Bickman, L. (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. Administration and Policy in Mental Health and Mental Health Services Research, 47, 795-843. https://doi.org/10.1007/s10488-020-01078-5
4. Chatterjee, M., Ghosh, S., & Chowdhury, S. R. (2021). Machine learning-based mental health prediction using survey data. Procedia Computer Science, 185, 293-302. https://doi.org/10.1016/j.procs.2021.05.031
5. Farooq, F., Khalid, H., & Khan, M. S. (2022). AI-assisted early diagnosis of bipolar disorder. Journal of Affective Disorders, 307, 224-233. https://doi.org/10.1016/j.jad.2022.04.029
6. Osborn, E., Wittkowski, A., Brooks, J., Briggs, P. E., & O'Brien, P. M. (2020). Women's experiences of receiving a diagnosis of premenstrual dysphoric disorder: a qualitative investigation. BMC Women's Health, 20(1). doi:10.1186/s12905-020-01100-8
7. Pim, C., Afzal, J., & Kamaldeep, B. (2023). The WHO World Mental Health Report: a call for action. The British Journal of Psychiatry, 222(6), 227-229. doi:10.1192/bjp.2023.9
8. Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research and Directions. Springer Link, 1-2. https://doi.org/10.1007/s42979-021-00592-x
9. World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization. Retrieved from https://www.who.int/publications/i/item/9789240049338
10. Zaagsma, G., Stökl Ben Ezra, D., Rürup, M., Margolis, M., & Levi, A. S. (2022). Jewish Studies in the Digital Age. Berlin, Boston: De Gruyter Oldenbourg. doi:10.1515/9783110744828