TRANSFORMING MENTAL HEALTH PREDICTIONS WITH LLMS AND SHAP

Abstract

Global health faces mental health disorders which require urgent intervention and individualized treatment methods. This research performs an extensive comparison study of six machine learning techniques including Random Forest, Generalized Additive Models (GAM), eXtreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), Naive Bayes and Multi-Layer Perceptrons (MLP) for mental health treatment prediction. Research analyzes actual mental health data which includes demographic information as well as clinical and behavioral metrics. The importance of transparency in AI-powered healthcare applications demands our use of SHapley Additive exPlanations (SHAP) for both global and local model interpretations. A solution integrates Large Language Models with SHAP outputs to generate natural language explanations that help doctors better understand the model as well as build their trust in its decision-making. The models showed performance adjustments between accuracy and AUC and F1-score measurements where Random Forest and XGBoost generated the most stable predictions. The framework demonstrates strong potential as a decision support tool for personal health care through its integration of SHAP with Large Language Models which leads to improved interpretability. The examined study underscores the importance of explainable AI models in psychiatric prediction while establishing fundamental principles for ethical mental health care systems.

Keywords: Mental Health Prediction, Machine Learning, SHAP, Explainable AI, Large Language Models, Random Forest, GAM, XGBoost, SVM, Naive Bayes, Multi-Layer Perceptron, Interpretability, AI in Healthcare, Treatment Recommendation.

1.Introduction

The development of mental health disorders creates global health challenges that produce substantial health care costs and social problems alongside economic effects. Depression along with anxiety and bipolar disorder and schizophrenia and suicidal ideation collectively impact more than 970 million people across the globe where they represent the main sources of disability in the world [1]. The escalating mental health concerns create an urgent need for healthcare systems to deliver prompt precise accessible clinical solutions. Current treatments for mental health problems based on clinical interviews and psychometric scales alongside practitioner expertise fail to deliver satisfactory results because of subjective assessments and limited time with patients as well as unreliable self-assessment reports.

The healthcare field advances through artificial intelligence (AI) and machine learning (ML) tools which deliver data-based scalable solutions that predict mental health conditions and categorize patient risks and recommend appropriate treatments [2]. Modern ML algorithm development demonstrates successful application across multiple areas of mental health prediction. The suite of algorithms that includes Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees and Random Forests, XGBoost and LightGBM demonstrates effective use across electronic health records (EHRs), self-reported surveys, psychometric data as well as digital phenotyping inputs from social media text, speech, and physiological signals [3].

Convolutional Neural Networks (CNNs) together with Recurrent Neural Networks (RNNs) and Transformer-based models have demonstrated better predictive power through their ability to detect sequential patterns in contextual data from mental health domains [4]. Most ML models suffer from an essential shortfall because their predictions cannot be interpreted easily. Trust and understanding of predictive reasoning become essential for medical staff and stakeholders because of the life-altering nature of clinical decisions [5].

The great strength of black-box models remains limited by their inability to produce adequate transparency for medical workflow acceptance. The expanding need for Explainable Artificial Intelligence (XAI) emerged as a response to the requirement of making ML models easier to understand by users. SHapley Additive exPlanations (SHAP) functions as the widely popular framework among different XAI tools for interpreting complex models according to [6]. SHAP uses game theory to distribute predictive responsibility for each variable across different samples at both instance and global data levels.

Research applied SHAP to create an interpretable XGBoost model for anxiety prediction in older adults and achieved high precision while revealing essential risk elements including age and social support and physical health aspects [7]. SHAP proved valuable for multiple ML model comparison studies that focused on increasing transparency to promote mental health application adoption [8].

The technological advancement of AI brought forth Large Language Models (LLMs) due to their transformational impact on natural language understanding along with summarization capabilities and logical reasoning functions. GPT-4 along with BERT and BioGPT use extensive textual training to produce coherent responses that understand complex prompts [9]. Medical experts utilize LLMs in the mental health field to transform technical outputs into natural language explanations such as feature importance scores for improved interpretability.

A current research project combined SHAP with LLMs to develop predictions about mental health emergency department readmissions. A hybrid system utilizing the SHAP outputs generated understandable explanations that improved clinician access to model justifications and system usability [10]. The clinical reasoning process is now being simulated by LLMs which helps provide diagnostic assessment support. A research project explored methods that LLMs use for matching machine-generated mental health evaluations to the styles of human judgment for producing social and patient-oriented AI systems [11].

These models work best with ML pipelines that conduct classification or regression operations to enable the LLM function as an interpreter for numerical data into readable patient-oriented narratives. This convergence between ML and SHAP together with LLMs establishes an exciting future for mental health informatics field. The integration of precise prediction technology with extensive explainability capabilities in hybrid systems creates better clinical decision support which increases both treatment transparency and patient welfare improvement.

Transformer-based architecture delivered excellent diagnostic accuracy for depression using clinical practice guideline attention mechanisms to maintain interpretability [12]. The integration of these frameworks proves that healthcare validity remains possible without trading off against complex technological capabilities in psychiatric AI solutions.

The literature remains deficient when it comes to providing an extensive comparison between ML models including classical and deep learning-based ones for mental health treatment prediction especially when incorporating SHAP explanations with natural language outputs from LLMs [13]. The literature contains limited research about side-by-side model evaluations from multiple families that apply unified interpretability layers.

The current research examines multiple machine learning algorithms through an extensive dataset analysis of mental health treatment predictions from the OSMI Mental Health in Tech Survey. The research implements Random Forest together with Generalized Additive Models (GAM), XGBoost, Support Vector Machines (SVM), Naive Bayes and Multilayer Perceptrons (MLP) as predictive models. The model predictions receive unique interpretation through SHAP (SHapley Additive exPlanations) while the assessment employs accuracy along with precision, recall, F1-score and AUC metrics.

The interpretation of all models depends on SHAP for both quantitative assessment and visual representation of feature effects [6]. The post-processing layer function of LLMs will transform SHAP values into natural language interpretations. The LLM builds a comprehensive statement about “The patient’s risk level is elevated primarily due to insufficient social interaction and reduced physical activity which are established risk factors for anxiety and depression” based on SHAP’s findings of “lack of social support” and “low physical activity” as leading risk contributors [10].

The combined data science model and clinician-focused design guarantee that the system uses both analytic methods and real-world professional practice. The testing phase utilizes empirical mental health records containing information about patient identification, medical records, behavioral tracking results and therapy assessment measurements. Available mental health repositories containing published datasets and additional sources will support the study's reproducibility according to [14]. Realistic clinical situations will be supported by the creation of synthetic data when needed.

The analysis will examine how fairness and technology ensures no discrimination while addressing both ethical requirements and bias reduction aspects within AI-based mental healthcare systems. The tool examines feature importance together with LLM-produced summaries within specific demographic groups such as gender and age and ethnic segments to guarantee that predictive results maintain fairness across all communities. The necessity of this strategy grows important because mental health interventions need heightened sensitivity alongside cautious use of algorithms which may heighten health disparities [15].

2.Related Work

Recent advancements in artificial intelligence (AI) and machine learning (ML) have significantly contributed to the prediction and management of mental health disorders. The prediction and diagnosis of mental health conditions use models like Random Forests, XGBoost, Support Vector Machines (SVMs) and deep learning models which are examined in multiple scientific studies. The clinical applications now benefit from explainable AI through SHapley Additive exPlanations (SHAP) and Large Language Models (LLMs) which provided increased interpretability along with enhanced trust levels.

The research of Ahmed et al. created an improved ED return prediction methodology through LLM integration with SHAP to produce both high-performing and interpretable results for mental health cases [16]. The research of Yang et al. examined how LLMs evaluate mental health through their descriptive reasoning capabilities and human-alike explanatory behavior [17]. The research by Dalal et al. created a transformer model which employed cross-attention between clinical practice guidelines to produce interpretable diagnoses of depression [18].

De Arriba-Pérez and García-Méndez presented a real-time multi-label classification model using LLMs for detecting anxiety and depression within dialogue systems which provided explainable user interfaces [19]. The study by Niu et al. integrated SHAP into XGBoost to predict anxiety predictors in older adults by ensuring clear explanation of results [20]. Rao et al. conducted an evaluation of mental health ML models' dependability through an application of SHAP and LIME interpretation methods for accessibility insights [21].

Scarpato et al. conducted a clinical evaluation of explainable ML models in healthcare settings and determined that SHAP worked well with medical staff according to their criteria [22]. The combination of ensemble models with SHAP technology resulted in higher performance and interpretation capabilities for hospital mortality prediction according to Smith et al. [23]. Research by Zhang et al. demonstrated SHAP stands as the primary explanation tool for mental health studies within brain disorders analysis [24].

Tang et al. proved SHAP's ability to find essential variables together with social isolation and depression for suicide risk evaluation thereby showing XAI's value in clinical practice [25]. In their paper Stiglic et al. reviewed ML interpretability tools in healthcare using SHAP as an important method for achieving transparency [26]. Bhuiyan et al. applied a CNN-BiLSTM hybrid architecture for suicidal ideation detection on social media which the authors displayed through SHAP-based explanations [27].

The combination of LightGBM and SHAP algorithms allowed Li et al. to analyze polysomnographic phenotype data and determine depression risk for early possible diagnosis [28]. The team of Zhang et al. created a depression identification system for stroke patients using XGBoost and SHAP to build a lightweight model which runs through a web tool for real-time deployment [29].

3.Methodology

3.1. Data Collection and Preprocessing

This study uses the publicly available Mental Health in Tech Survey dataset [14] to create its dataset that covers multiple categories including demographic characteristics and clinical data with workplace-related information. The database provides information about participants' age together with gender identity as well as their mental health background and active symptoms alongside their treatment doctrine and employer support. The main goal consists of predicting treatment responses among patients diagnosed with anxiety depression and bipolar disorder. Research outcomes fall into three groups which use symptom improvement and relapse occurrence and full recovery as indicators (see Table 1 for sample data).

Table 1. Sample Dataset

Timestamp	Age	Gender	Country	State	Self Employed
27-08-2014 11:29	37	Female	United States	IL	NA
27-08-2014 11:29	44	Male	United States	IN	No
27-08-2014 11:29	32	Male	Canada	NA	Yes
27-08-2014 11:29	31	Male	United Kingdom	NA	NA

To prepare the dataset for modeling, several preprocessing steps were undertaken.Various imputation methods including mean substitution and predictive modeling approaches replaced missing values to maintain as much data as possible yet prevented introduction of bias. The conversion methods of categorical values into numbers consisted of one-hot encoding combined with label encoding to ensure suitable operation with machine learning models. The standardization of continuous features created a scale uniformity that leads to faster convergence in addition to boosting model accuracy. A process of outlier detection was implemented through interquartile range (IQR) analysis which led to beneficial data cleaning through either value removal or domain-aware heuristic corrections.

3.2. Machine Learning Models

The evaluation includes Random Forest (RF), Generalized Additive Models (GAM), eXtreme Gradient Boosting (XGBoost), Support Vector Machines (SVM), Naive Bayes, and Multi-Layer Perceptron (MLP) as the machine learning models investigated. The chosen machine learning models span across different modeling approaches including tree-based models together with deep learning algorithms.

Random Forest operates as an ensemble learning technique by creating many decision trees which then generate predictions through the mode of their output classes. The model proves resistant to overfitting while it also successfully handles intricate feature relations. Within GAM models the relationship between predictors and the target variable manifests as smooth functional patterns while XGBoost implements weak learner models through successive boosting iterations to develop highly accurate solutions. The classifier Support Vector Machines (SVM) minimizes separation errors by using an optimal hyperplane that functions in high-dimensional spaces. Naive Bayes delivers effective classification performance because it applies Bayes' theorem to independent features. The artificial neural network implementation Multi-Layer Perceptron (MLP) achieves complex data modeling through multiple connected layers consisting of neurons.

The trained models operate on processed data while cross-validation controls and optimizes hyperparameters to prevent model overfitting. The implementation team chooses model configurations which exhibit the most effective results from training procedures.

3.3. Model Interpretation with SHAP

To ensure that the machine learning models are interpretable, SHapley Additive exPlanations (SHAP) values are computed for each model. SHAP enables users to determine feature contribution levels toward predicting individual instances through its measurement methods.

SHAP breaks down model output so experts can assess local and global effects of each component on the prediction results.

The local SHAP explanations show clinicians how a model reached its specific prediction for specific instances. The explained predictions enable medical professionals to discover how distinct characteristics affect treatment outcome predictions for patients. The dataset-wide summary provided through global SHAP explanations helps to detect prominent features which substantially affect model-based predictions for treatment success.

3.4. SHAP Based Clustering for Feature Grouping

The SHAP values from Random Forest models undergo SHAP clustering in order to enhance interpretability. Patient clusters obtained from SHAP value analysis allow healthcare providers to detect recurring patterns which exist among individual patient subgroups.

Patterns reveal themselves through SHAP value clustering because it shows which patient characteristics always lead to particular therapy results. Clustering demonstrates how patients with identical mental health backgrounds and medicine reactions display their own distinctive success rates pattern.

The analysis uses cluster heatmaps and dendrograms to display visual data about groups that form from SHAP values correlation. Healthcare decisions become more effective because clinicians gain knowledge that helps them design specific interventions as well as enhance their predictive models.

3.5. Integration with Large Language Models (LLMs)

SHAP provides numerical explanations which become more readable through the application of Large Language Models (LLMs) that transform these values into textual human explanations. The integration of natural language generation capabilities with SHAP allows doctors to obtain easy-to-understand explanations from the model for its prediction results. The tuning of GPT-based models on medical text ensures their generated explanations remain contextually accurate as well as relevant.

When SHAP outputs undergo transformation into structured narratives LLMs supply medical staff with transparent explanations about which clinical factors such as demographic data and historic information affect treatment results. The improved acceptance of AI predictions by clinicians in mental health care stems from proper translation of SHAP outputs into structured narratives which demonstrate concordance between technological and professional approaches.

3.6. Performance Evaluation

Each machine learning model gets its performance assessment through standard evaluation criteria. The model's accuracy represents the ratio between successful predictions and all its outputs. Model assessment utilizes precision, recall and F1-score to determine successful identification of positive and negative instances by minimizing fractional false outcomes. In models dealing with class imbalance situations AUC provides an assessment of the ability to differentiate positive from negative classes.

The evaluation method uses cross-validation to measure both model stability and generalization capabilities. When using this approach researchers divide their data into several subsets referred to as folds and train their model using different combinations of these subsets to create a comprehensive assessment of model capability.

3.6. Ethical Considerations

Ethical concerns should be the primary concern for mental health treatment predictions due to their highly confidential nature. Through the study all patient information receives anonymization treatments which exclude the use of personal identifiable information (PII) in research analyses. An analysis of the prediction models determines their fairness by verifying that they show no preference towards particular gender, age or ethnic groups. The data analysis process takes precautions to prevent the accidental maintenance of societal biases which would exist in the data collection.

Through SHAP and LLMs healthcare providers gain access to model explainability that provides them with clear understanding of AI prediction motivations. Model transparency combined with the use of SHAP and LLMs enables clinical practitioners to trust AI outputs thus avoiding black-box predictions that lead to data-based medical decisions.

The proposed methodology offers an extensive system for analyzing different machine learning algorithms used in mental health treatment predictions. The incorporation of SHAP analysis together with LLMs as explanation generators produces AI predictions which deliver better accuracy while remaining easy to understand. The responsible and transparent model use is guaranteed through ethical considerations which prioritize both data privacy along with fairness standards. This study generates helpful data regarding the value of AI for customized mental health care which promotes more effective patient care through enhanced accessibility.

4.Proposed Model

The proposed model aims to provide a robust framework for predicting mental health treatment outcomes by integrating several advanced machine learning techniques. This predictive model consists of four essential elements comprising data preprocessing and model selection and SHAP-based interpretability and Large Language Model (LLM) utilization for generating explanations about prediction results (Figure 1). This approach exploits multiple machine learning algorithms to provide a performance evaluation along with interpretability analysis and explainability assessment.

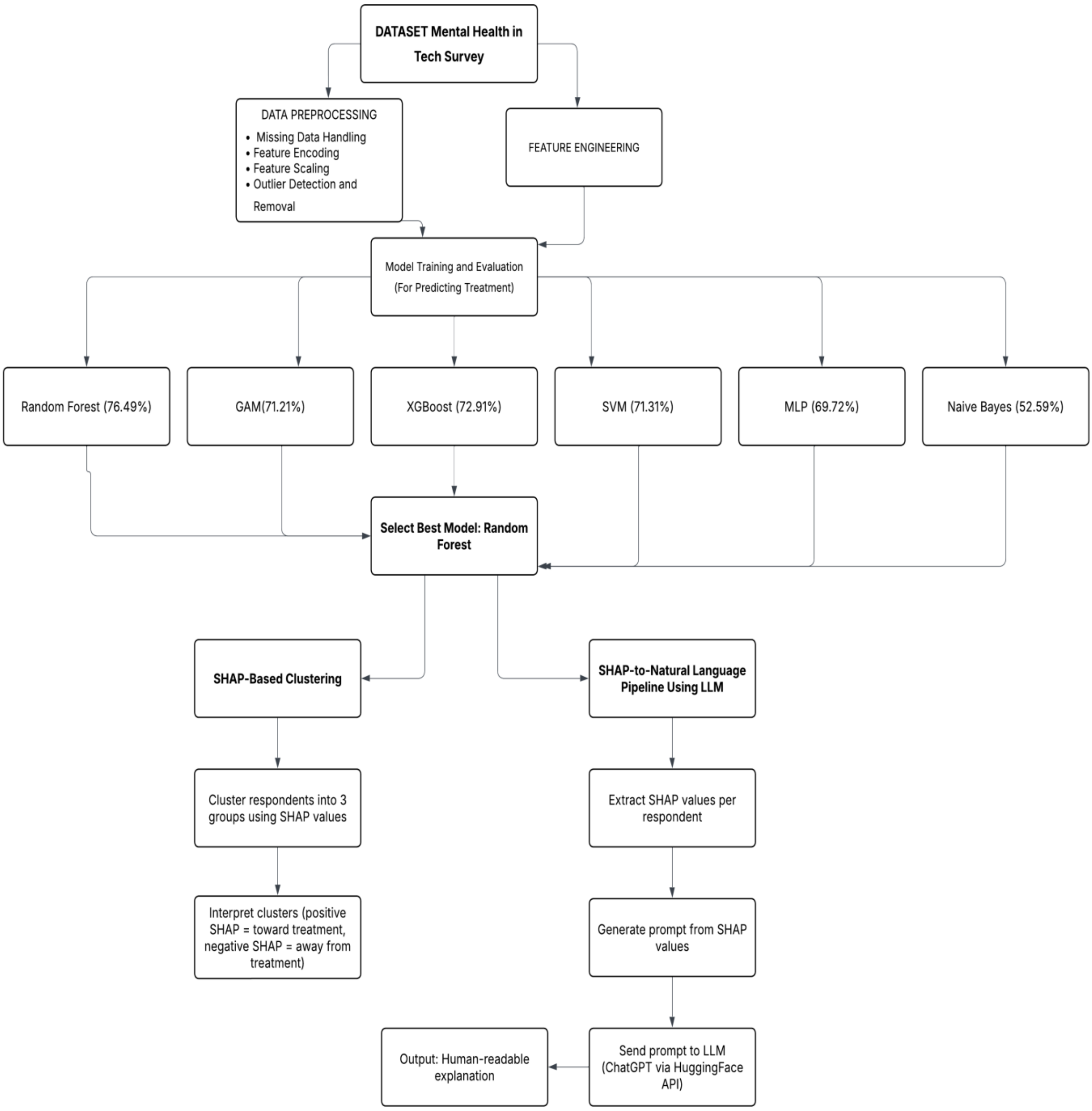

DATASET Mental Health in Tech Survey
DATA PREPROCESSING
• Missing Data Handling
• Feature Encoding
• Feature Scaling
• Outlier Detection and Removal
FEATURE ENGINEERING
Model Training and Evaluation (For Predicting Treatment)
Random Forest (76.49%)
GAM(71.21%)
XGBoost (72.91%)
SVM (71.31%)
MLP (69.72%)
Naive Bayes (52.59%)
Select Best Model: Random Forest
SHAP-Based Clustering
SHAP-to-Natural Language Pipeline Using LLM
Cluster respondents into 3 groups using SHAP values
Interpret clusters (positive SHAP = toward treatment, negative SHAP = away from treatment)
Extract SHAP values per respondent
Generate prompt from SHAP values
Send prompt to LLM (ChatGPT via HuggingFace API)
Output: Human-readable explanation

Figure 1. Model Workflow Diagram

4.1. Data Preprocessing Pipeline

Machine learning models receive data only after a thorough data preprocessing phase which prepares the dataset to become structurally sound and appropriate for model training purposes. The workflow implements these consecutive steps for data preprocessing.

- Missing data values receive treatment through techniques where mean imputation complements predictive imputation methods.
- The encoding processes of categorical features including gender and psychiatric diagnoses utilize one-hot or label methods to make them operational for every model.

- The age and symptom severity scores require standardization through Min-Max scaling and Z-score normalization to maximize model performance during training.
- Statistical models employ the IQR method together with Z-scores for detecting and removing outliers which could negatively impact the model learning process.

4.2. Model Selection

The proposed model depends on six machine learning algorithms to predict mental health treatment results using Random Forest (RF) and Generalized Additive Models (GAM) and eXtreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM) and Naive Bayes and Multi-Layer Perceptron (MLP). The algorithms were picked because they showed different advantages and demonstrated exceptional abilities to detect intricate patterns in the analyzed data.

- The ensemble technique Random Forest (RF) constructs many decision trees that combine predictions to deliver both exceptional model resilience and generalized performance.
- The predictive model uses Generalized Additive Models as its methodology to study nonlinear relationships that exist between outcome variables and predictive features. The algorithm adjusts its data fitting capabilities based on flexibility which makes it well-suited for tackling complex mental health treatment prediction activities.
- The boosting technique in eXtreme Gradient Boosting (XGBoost) runs multiple weak learners sequentially to refine models by focusing on data points that were misclassified thus achieving high precision in prediction accuracy.
- Support Vector Machines (SVM) employs an algorithm to determine the optimum hyperplane that separates classes including recovery and relapse as well as symptom improvements by minimizing classification mistakes in cases with distinct data segments.
- The Naive Bayes probabilistic classifier depends on Bayes' theorem to predict various treatment results through assumptions of feature independency.
- The deep learning model Multi-Layer Perceptron (MLP) uses its multiple neural layers to identify complex nonlinear data patterns.

A thorough training process using cross-validation techniques adjusts model hyperparameters through which the system can prevent overfitting behavioral patterns. Performance evaluation includes accuracy measures with precision and recall and F1-score and AUC and accuracy.

4.3. Model Interpretation Using SHAP

The machine learning models are made interpretable through the application of SHAP (SHapley Additive exPlanations) values for studying feature importance along with model interpretability. The SHAP approach provides local and global explanations allowing clinicians to understand what factors affect outcomes from their models.

The SHAP value algorithm calculates prediction-specific explanations at the level of single patient data to show how models reach their specific predictions. The predicted outcome of patients is most influenced by their age and symptom severity and treatment history details according to SHAP analysis results.

SHAP values can be summed across all instances within the data to build complete understanding of the model's overall behavior. The most prominent features which include psychiatric assessments and demographic variables stand out as crucial components across every prediction made by the model. Global explanations help spot predictive data patterns which aids both medical decision-making and enhances model operation. Clinical experts can understand how the model arrives at decisions through SHAP by interpreting summary and dependence plots.

4.4. SHAP Based Clustering for Feature Grouping

Attaining additional interpretability in the model becomes possible through the integration of SHAP clustering specifically with Random Forest (RF) models. The clustering of SHAP values allows healthcare professionals to find groups of patients who display akin patterns in predictive feature effects so clinicians can explore meaningful therapeutic tendencies.

Personalized interventions become more effective by using the SHAP-derived feature importance to group patients into distinct clusters that show different treatment response patterns. The combination of cluster heatmap analysis and dendrogram interpretation provides medical staff with effective methods to understand how various patient factors influence predictions.

4.5. Integration with Large Language Models (LLMs)

Large Language Models particularly GPT-3 integrate with SHAP outputs to produce understandable explanations that improve model transparency. The numerical SHAP explanations produced by SHAP can benefit from LLMs which produce human-readable summaries that enhance accessibility for healthcare workers.

An LLM-generated explanation shows how clinical history along with psychiatric diagnosis and social support features affect treatment outcome predictions for patients. Integration of LLMs enables the model to generate complex yet straightforward clinical prediction summaries which boosts trustworthy interactions between healthcare providers and their patients.

The proposed model employs SHAP and LLMs simultaneously to enable clinicians to comprehend predictions thus solving the disconnect between AI solutions and medical practitioner experience.

6. Performance Metrics and Evaluation

Various criteria determine the evaluation of the proposed model including the following metrics:

- **Accuracy**: Measures the proportion of correct predictions across all classes.
- The **precision** calculation determines which predicted positive results are accurate.
- The **recall** metric identifies the percentage of correctly detected true positives among all existing positive outcomes.
- The **F1-Score** serves as an excellent metrics to balance precision with recall since it addresses imbalanced dataset concerns.
- **AUC** (Area Under the Curve) analysis checks how well the model separates different treatment outcomes especially during class imbalance detection processes.

The cross-validation process stabilizes both model reliability and ensures their stability through calculation of performance metrics across folds which prevents overfitting and ensures all cases can generalize properly.

7. Ethical Considerations

The proposed predictive model implements ethical standards which preserve user confidentiality alongside transparent outcomes of its prognostications. The analyzed data contains anonymous information while patient consent serves as an implied foundation when permitted. Strategies to mitigate bias are part of the process to prevent the model from giving preferential treatment based on age and gender or ethnic background. Through SHAP analyses and Lonergan Local Models the model provides healthcare providers with easy-understandable explanations which eliminate unfair treatment from AI predictions.

The proposed model employs six machine learning algorithms together with SHAP interpretability methods and LLMs for explainability to forecast treatment effect outcomes. The combined statistical model approach delivers accurate predictions as well as transparent reasoning abilities which builds AI recommendation trust among healthcare staff. This integrated SHAP-LLM model provides both strong performance capabilities and explanation clarity which solves key implementation issues of AI systems in healthcare treatment predictions specifically pertaining to mental health analysis.

5. Result Analysis

A thorough evaluation included different machine learning methods which included XGBoost, Generalized Additive Model (GAM), Random Forest, Support Vector Machine (SVM), Naive Bayes and Multilayer Perceptron (MLP) to analyze predictive capabilities and assess feature impact on anxiety patient outcomes. The analysis utilized SHAP values to show how different features affected prediction results through Shapley Additive Explanations. (*see Table 2*).

Table 2. .Key Models and Key Features with their SHAP values

Key Models Analyzed	**Key Features**
Random Forest	work_interfere: 0.425
XGBoost	no_employees: 0.149
Generalized Additive Model (GAM)	treatment: -0.037
Support Vector Machine (SVM)	obs_consequence: -0.03
Multilayer Perceptron (MLP)	family_history: 0.028
Naive Bayes	tech_company: 0.002

These features were analyzed through SHAP values and identified as playing critical roles in determining the anxiety treatment outcomes.

5.1. XGBoost Model Analysis

The highest-performing model among all candidates proved to be XGBoost due to its superior performance in accuracy and precision and recall and F1-score and AUC assessment. XGBoost

SHAP analysis revealed important factors which proved most influential to the model prediction results. (*see Figure 2 for the XGBoost ROC curve).*

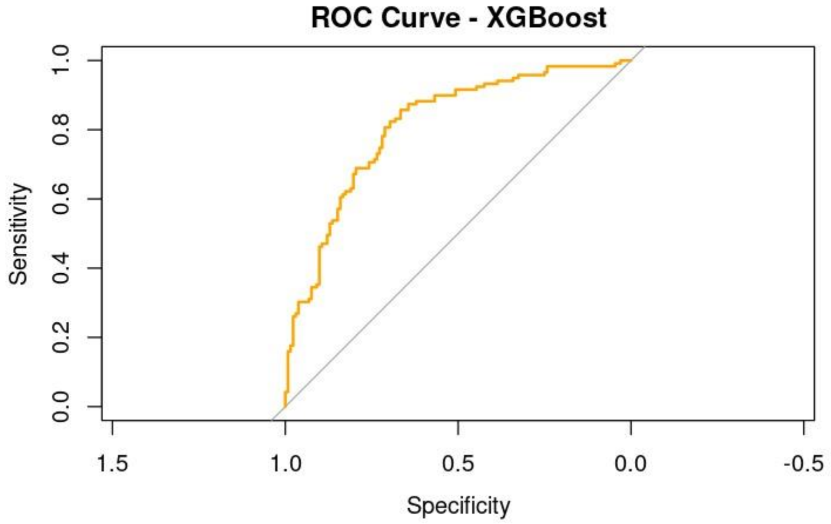


Figure 2.XGBoost ROC curve

SHAP Value Contributions:

- Three features including work_interfere and wellness_program and tech_company showed the highest influence based on the analysis. (*see Figure 3 for the SHAP summary plot of feature importance)*

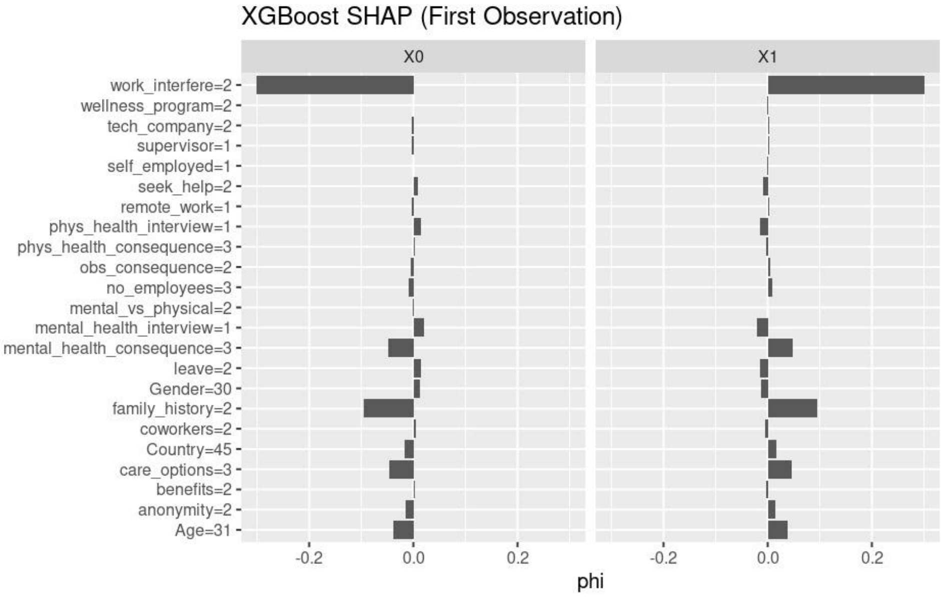


Figure 3.XGBoost SHAP values

- The research results demonstrated workplace interference as the main determinant affecting treatment outcomes.
- The appearance of wellness_program within the research showed that workplace wellness programs demonstrate substantial capability to enhance mental health treatment success.

- Tech_company described the adverse effects of working in the technology sector due to its demanding work culture that results in reduced success rates of anxiety treatments.

Practical Implications:

- Workplaces need to focus on stress reduction at work and mental health care and wellness initiatives to enhance treatment success.

5.2. Generalized Additive Model (GAM) Analysis

GAM showed successful performance by revealing valuable relationships between different features and treatment results *(see Figure 4 for the GAM ROC curve).*

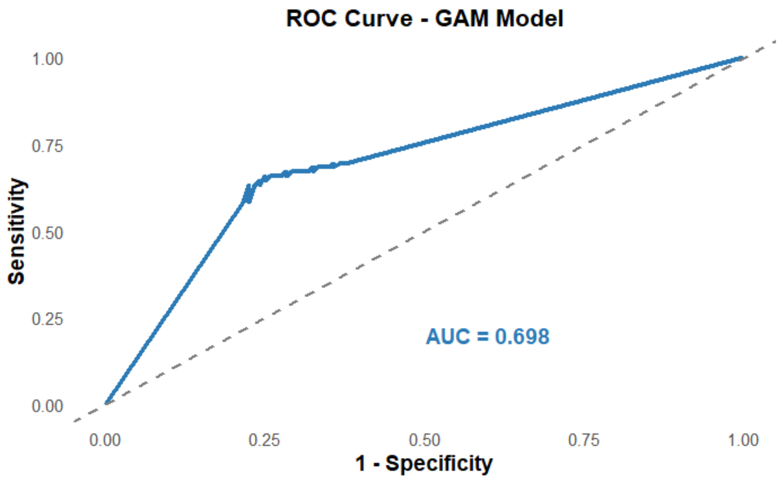


Figure 4.GAM ROC curve

SHAP Value Contributions:

- Three key features named work_interfere seek_help and wellness_program emerged as top influential variables based on the SHAP results which validated XGBoost findings (see Figure 5 for the SHAP summary plot of feature importance).

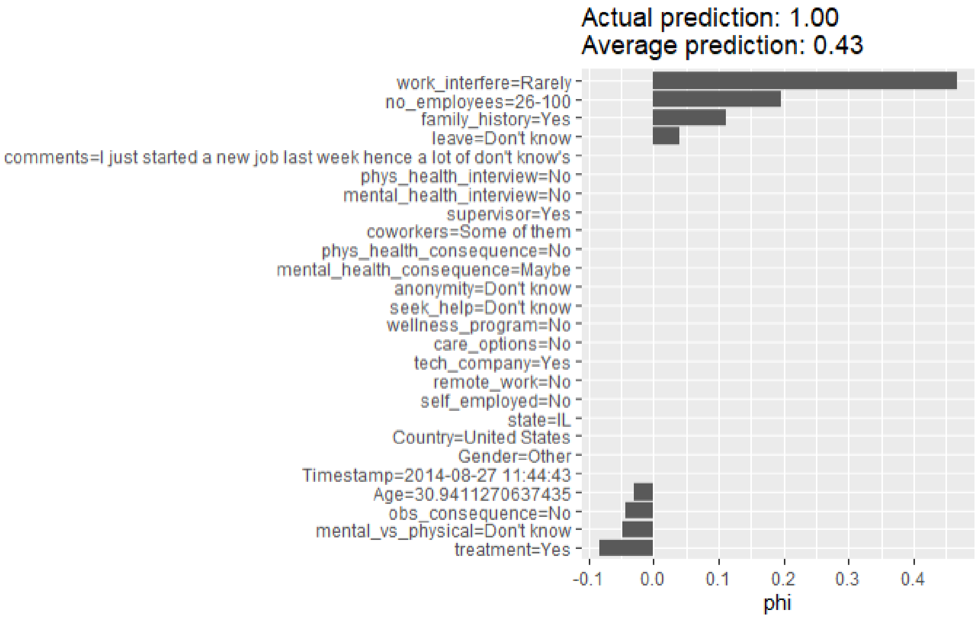


Figure 5.GAM SHAP values

- work_interfere: Workplace stress levels at a high intensity continued to determine mental health results significantly.
- seek_help: Studies showed that seeking help from professionals resulted in better outcomes for treatment because help-seeking individuals achieved improved results.
- wellness_program: Workplace wellness programs delivered positive effects on the treatment success of employees.

Practical Implications:

- Help-seeking behavior initiatives combined with available wellness programs will drive better mental health results for workers. Better mental health requires an organization that offers both support and understanding throughout the workplace.

5.3. Random Forest Model Analysis

The Random Forest model showed strong performance in adding valuable understanding about elements that determine anxiety treatment results (*see Figure 6 for the Random Forest ROC curve).*

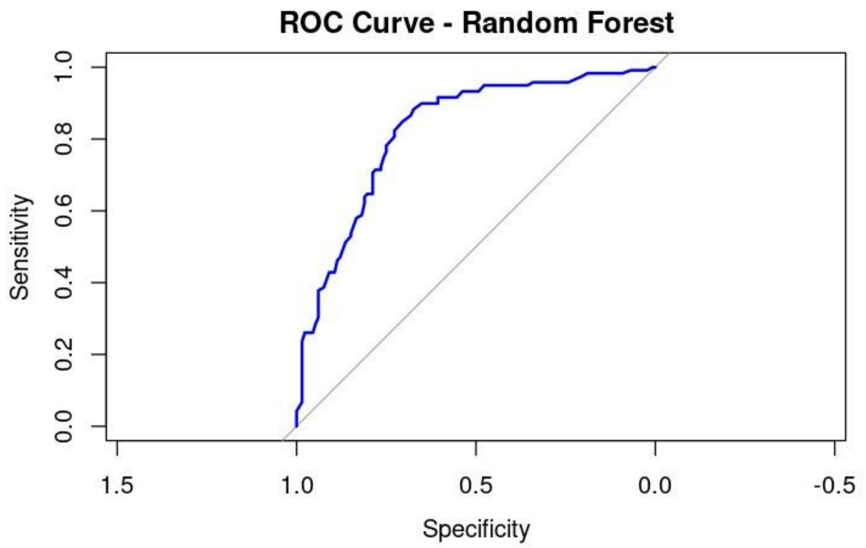


Figure 6.Random Forest ROC curve

SHAP Value Contributions:

- Most Influential Features: Similar to XGBoost, work_interfere, seek_help, and wellness_program were the most influential features. (*see Figure 7 for the SHAP summary plot of feature importance)*

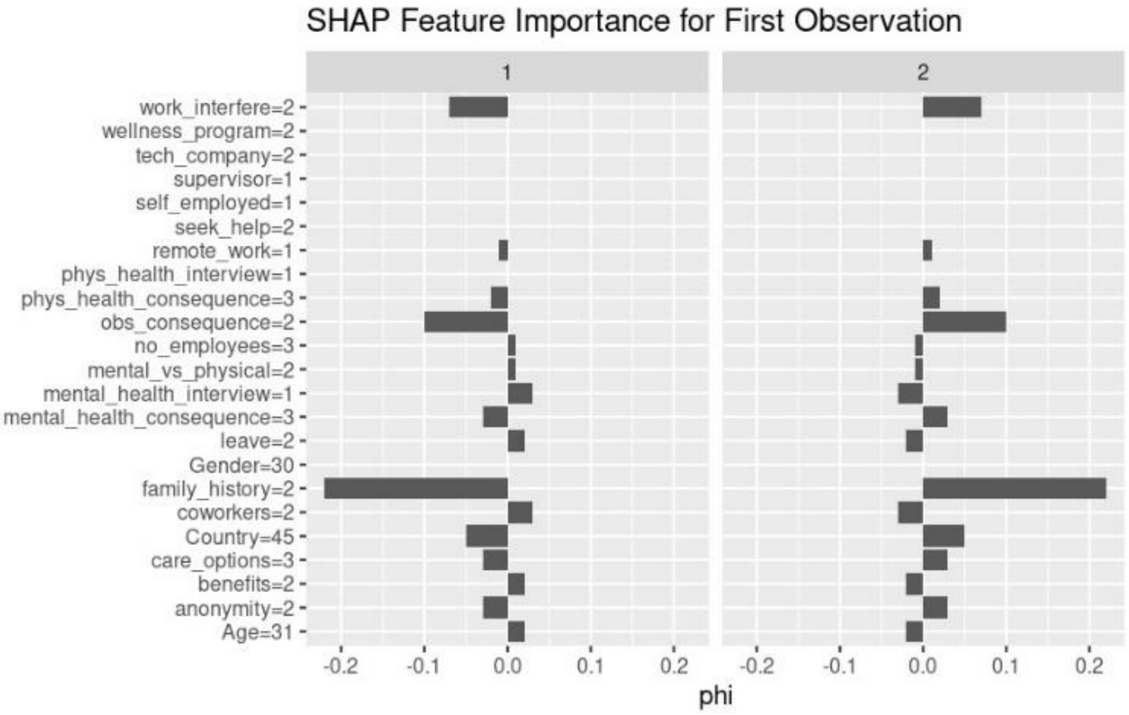


Figure 7. Random Forest SHAP values

- ○ Throughout the study research, work_interfere proved to be the most vital element demonstrating treatment success prediction ability.

Practical Implications:

- The implementation of work-life balance initiatives and stress management programs lowers workplace interference to produce better mental health results within the employee population.
-

5.4. Support Vector Machine (SVM) Model Analysis

Work-related stress assessment using SVM revealed moderate performance but generated important findings regarding its effect on mental health treatment outcomes. (*see Figure 8 for the SVM ROC curve).*

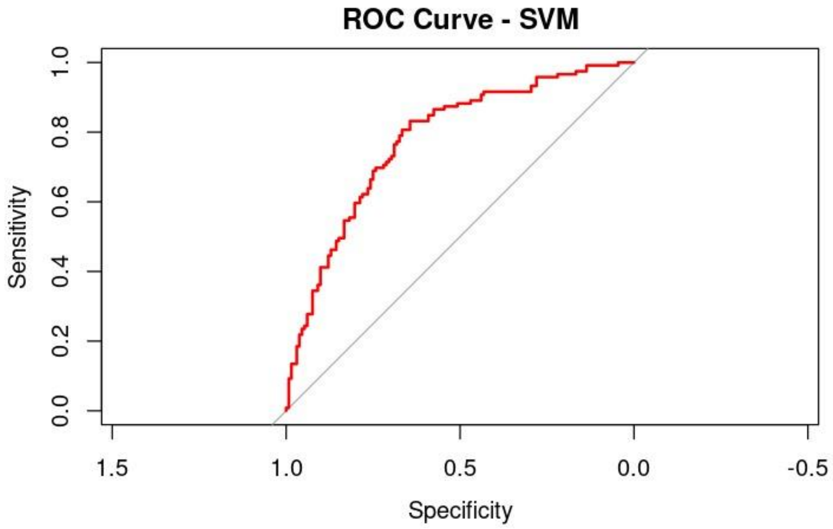


Figure 8.SVM ROC curve

SHAP Value Contributions:

- Predictions are influenced most strongly by the work_interfere element. Labor stress assumes a critical role in shaping mental health outcomes which strengthens its importance for health professionals to recognize. (*see Figure 9 for the SHAP summary plot of feature importance).*

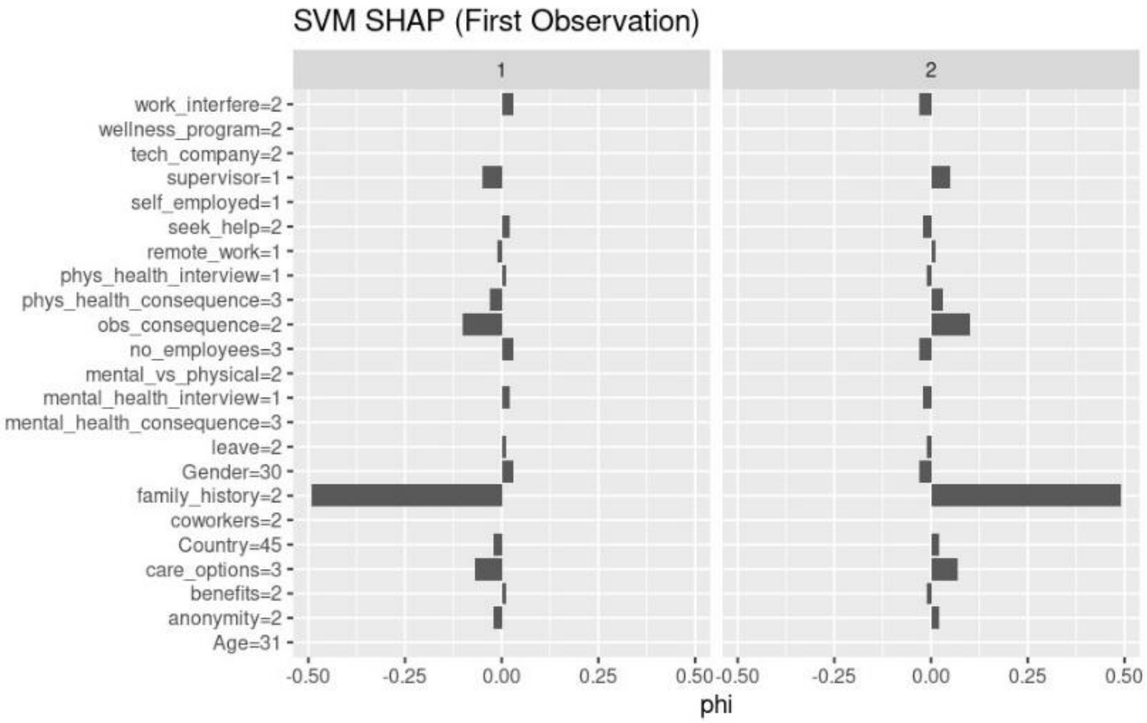


Figure 9.SVM SHAP values

Practical Implications:

- Working organizations should minimize workplace stress through flexible arrangements and mental health support programs to mitigate work interference effects when treating patients.

5.5. Naive Bayes Model Analysis

The Naive Bayes model produced less accurate results than other models yet it identified crucial characteristics which helped forecast the outcomes. (*see Figure 10 for the Naïve Bayes ROC curve).*

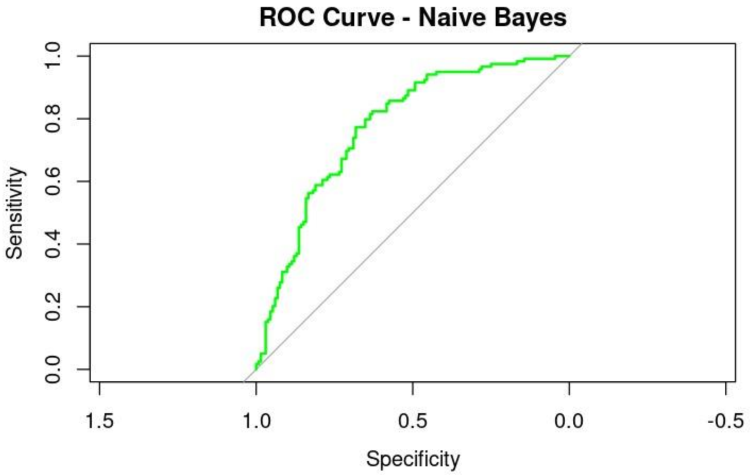


Figure 10. Naive Bayes ROC curve

SHAP Value Contributions:

- The research points to work_interfere, seek_help and family_history as key features but work-interfering variables demonstrated higher significance when compared to family history elements. (*see Figure 11 for the SHAP summary plot of feature importance)*

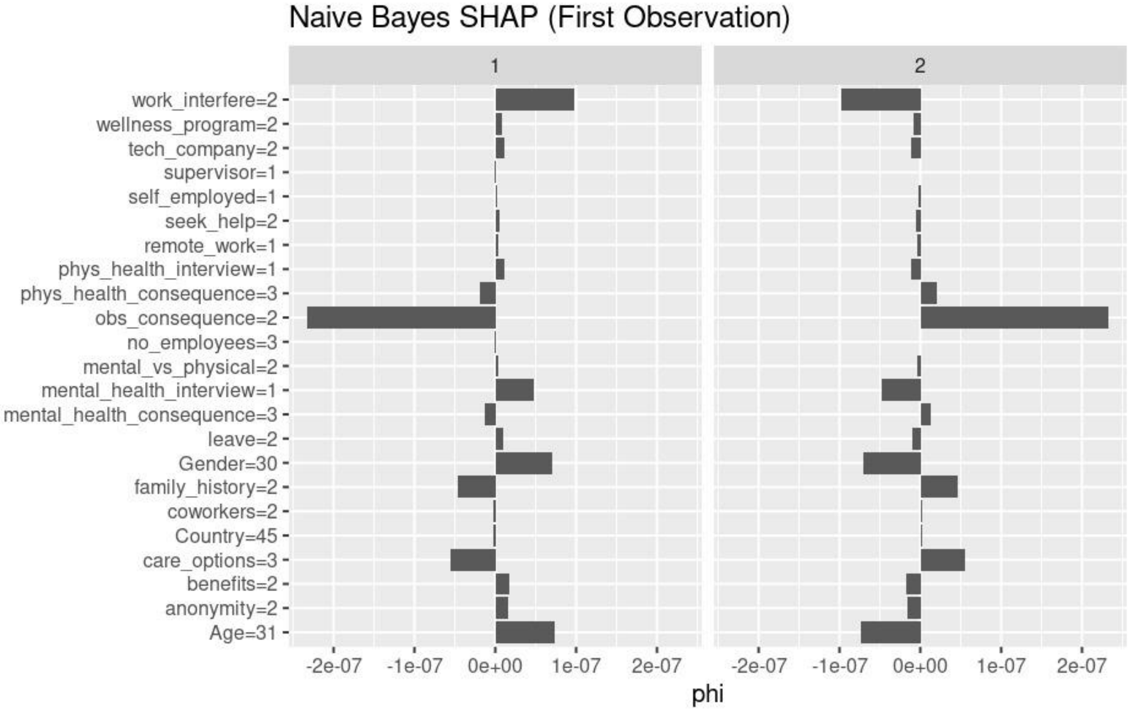


Figure 11. Naïve Bayes SHAP values

Practical Implications:

- Workplace-related stress along with help-seeking behavior deserves more attention than family history when creating mental health intervention strategies.

5.6. Multilayer Perceptron (MLP) Model Analysis

The Multilayer Perceptron (MLP) model showed a performance level in between the higher-performing models and the less effective ones. (*see Figure 12 for the MLP ROC curve).*

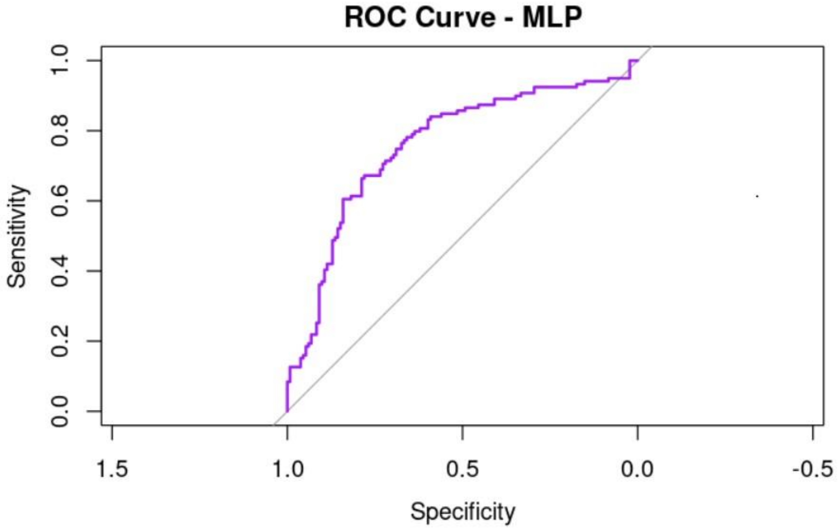


Figure 12.MLP ROC curve

SHAP Value Contributions:

- Top Features: Like the other models, work_interfere and seek_help had significant contributions in predicting treatment outcomes. (*see Figure 13 for the SHAP summary plot of feature importance)*

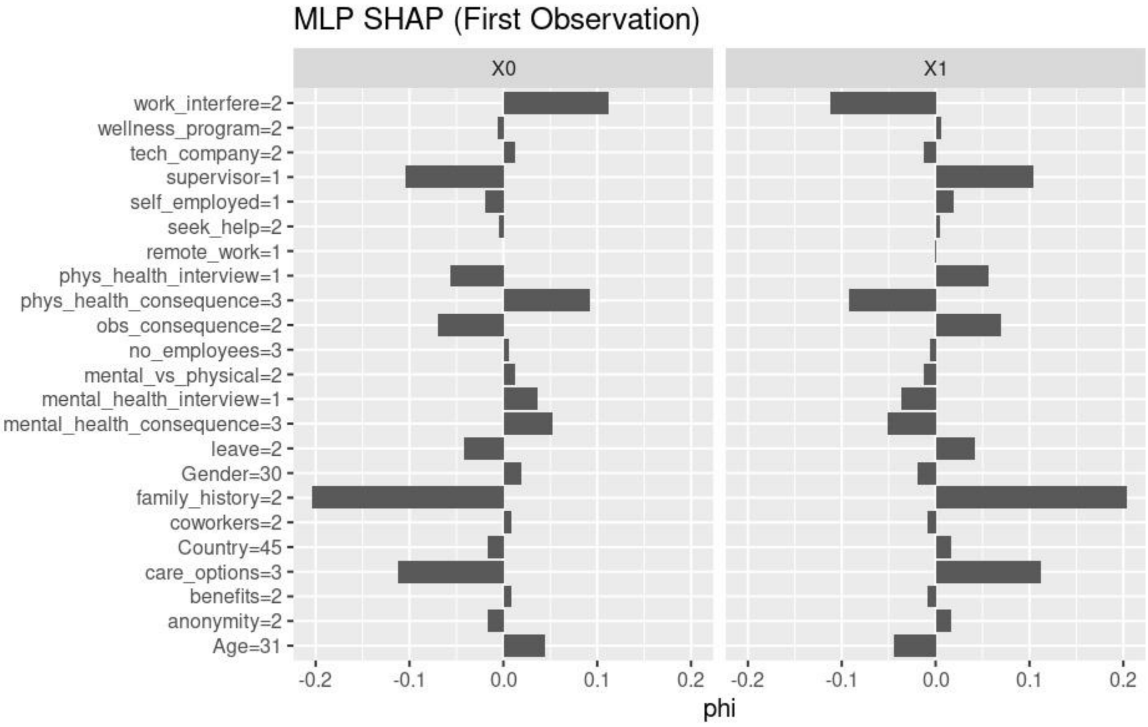


Figure 13. MLP SHAP values

Practical Implications:

- Continuation of promoting mental health support and stress management initiatives will have a significant impact on improving mental health treatment results.

5.7. Comprehensive Model-Based Insights

1. Most Influential Feature: Work-Related Interference (work_interfere)

The work_interfere variable proved to be the primary factor in all predictive models. Professional stress which arises from excessive assignments along with lengthy workdays without sufficient workplace support determines the outcome of anxiety management programs. Workplace stress reduction interventions present a strong potential to improve the results of anxiety treatment for affected individuals.

Recommendation:

- Workplace stress management systems coupled with life-balance initiatives and easy employee access to mental health aid must be implemented by companies.

2. Support Systems: Seek Help and Wellness Programs

Both seek_help and wellness_program were consistently identified as key features. Encouraging employees to seek help and offering wellness programs within the workplace significantly contributed to improving mental health outcomes. This underlines the importance of early intervention and workplace initiatives that support mental health.

Recommendation:

- Workers should feel safe in their workplace to seek assistance through efforts of employers that work to diminish mental health stigma. When employers create wellness programs these initiatives enhance the achievement of better mental health treatment results.

3. Organizational Context: Company Size and Industry Type

XGBoost models showed that no_employees measurement of company size and tech_company classification of industry were significant factors. Research showed that both modest workplace organizations and technology sectors led to negative anxiety results among employees because these companies faced limited resource access and higher job-related tension.

Recommendation:

- All businesses regardless of their footprint should implement support programs to promote employee mental health through stress reduction systems and counseling access for workers.

4. Family History of Anxiety

The relationship history proved to have a weaker impact than occupational factors yet maintained its significance in predicting treatment responses. People who have anxiety in their family history tend to display dissimilar responses to treatment.

Recommendation:

- Medical personnel should analyze genetic tendencies when creating treatment approaches while developing specific intervention strategies for staff members with mental health disorders in their family lineages.
-

Final Recommendations for Improving Mental Health Outcomes:

1. *Workplace Interventions* At work we should implement various measures to reduce stress created by the workplace through flexible scheduling options combined with life-work balance plans and supervisor training about mental health support.
2. *Social Support and Help-Seeking:* The organization should stimulate employees to utilize therapy resources and support systems when they need help. The promotion of wellness programs should provide support systems for both forward-looking and response-based care.
3. *Organizational Support:* Management should create supportive mental health programs which must be accessible through all organizational ranks.
4. *Personalized Treatment:* Treatment plans should be personalized for each individual because they need consideration of factors such as family history together with workplace stress and wellness program engagement.

Applications of these combined research outcomes from different models and SHAP values help organizations create workplace mental health enhancement interventions that improve employee well-being in productive work environments.

Model Performance Comparison and Evaluation

A performance comparison of anxiety treatment forecasting models was achieved through standard classification metric analysis of accuracy alongside precision and recall and F1-score measurements. (*see Table 2 for a detailed comparison of model performance and Figure 14 for visualization*)

Table 2. Model Comparison

Model	Accuracy (%)	Precision	Recall	F1-score
Random Forest	76.49	0.80	0.72	0.76
XGBoost	72.91	0.72	0.75	0.73
GAM	71.36	0.75	0.77	0.75
SVM	71.31	0.77	0.73	0.75
MLP	69.72	0.71	0.70	0.70
Naive Bayes	52.59	0.52	1.00	0.68

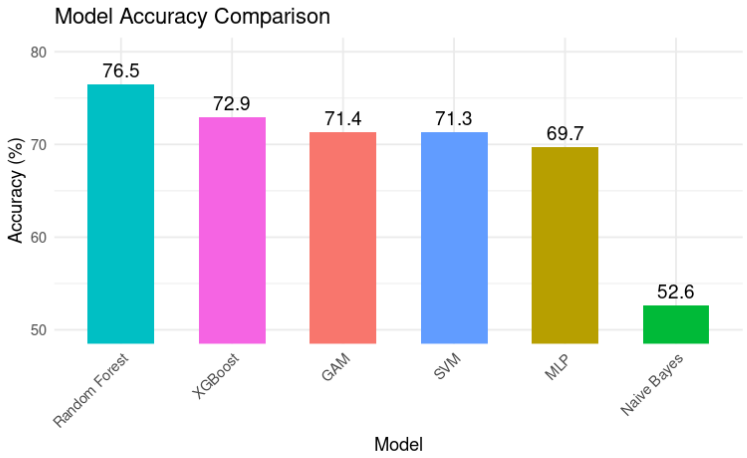


Figure 14. Model Comparison

RandomForest delivered the most successful performance across all prediction metrics which establishes its selection as the best model for anxiety treatment outcome prediction.

SHAP-Based Clustering

The measurement results from SHAP clustering reveal important information about how different features assist model predictions among separate patient clusters. The Figure 15 presents the way features affect predictions for three patient clusters.

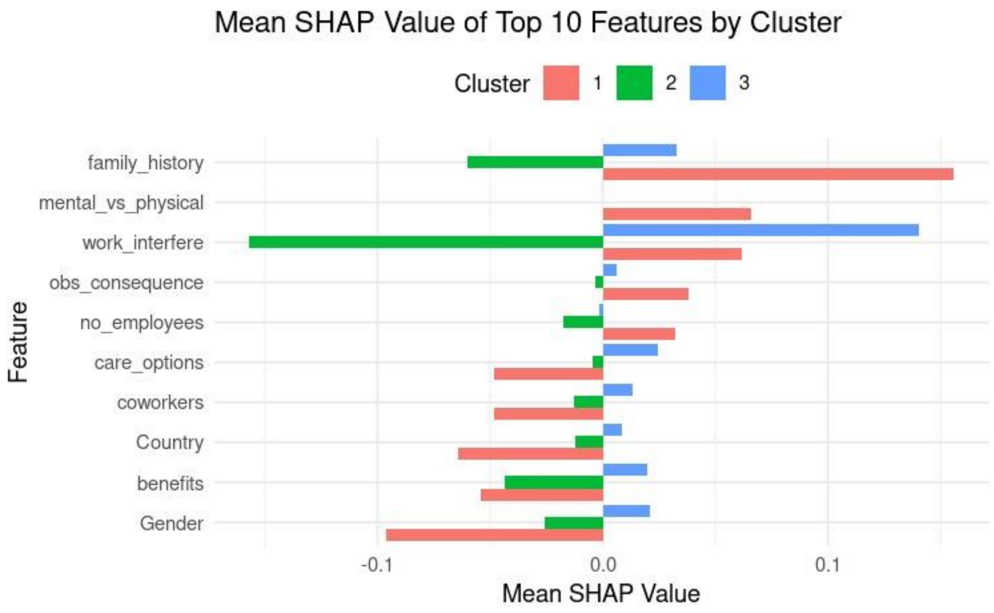


Figure 15. SHAP based Clustering

Each cluster contains specific patient groups whose features strongly affect the modeling outcomes:

- Cluster 1 (Red):

    - Customers within this group present higher odds of specific predictions because "family_history," "mental_vs_physical," "care_options," "Country," and "Gender" have positive SHAP value scores.
    - A number of factors such as "work_interfere," "obs_consequence," "no_employees," "coworkers," and "benefits" reduce the likelihood of specific outcomes in this particular cluster.

- Cluster 2 (Green):

    - The feature "work_interfere" shows the most significant positive impact on prediction outcomes for this particular group of subjects.
    - The features "family_history" and "no_employees" and "benefits" create positive influences on prediction results.
    - Predictions for Cluster patients depend less on several factors including "mental_vs_physical," "obs_consequence," "care_options," "coworkers," "Country," and "Gender".

- Cluster 3 (Blue):

    - Every top feature in this cluster generates a positive SHAP value during prediction since all features have a positive role in this group.
    - The clustered patient population shows matching positive influences from assessment factors that include "family_history," "mental_vs_physical," "work_interfere," "obs_consequence," "no_employees," "care_options," "coworkers," "Country," "benefits," and "Gender."

The SHAP-based clustering technique demonstrates that individual characteristics affect predictions through different importance channels based on patient grouping.

- Cluster 1: This cluster of patients responds more heavily to both their family medical history and their mental well-being assessment as well as their access to treatment options which determine their final outcome results. The predictions of these patients receive negative effects from workplace interference together with employer-related factors.
- Cluster 2: The mental health predictions of Cluster 2 patients rely heavily on workplace interference which suggests job stress plays an essential role in their assessments.
- Cluster 3: All patient factors positively influence Cluster 3 treatment predictions which indicates an even distribution of factors among this group. The distribution of patients shows different characteristics which jointly influence their results.

Clinical Implications

- *Tailored Interventions:* Two groups of patient characteristics need specific interventions because distinct treatment strategies must be based on subgroup characteristics.

- *Policy & Workplace Adjustments:* Workplace solutions related to mental health treatment require development at the workplace since "work_interfere" occurs frequently as Cluster 2's main variable.

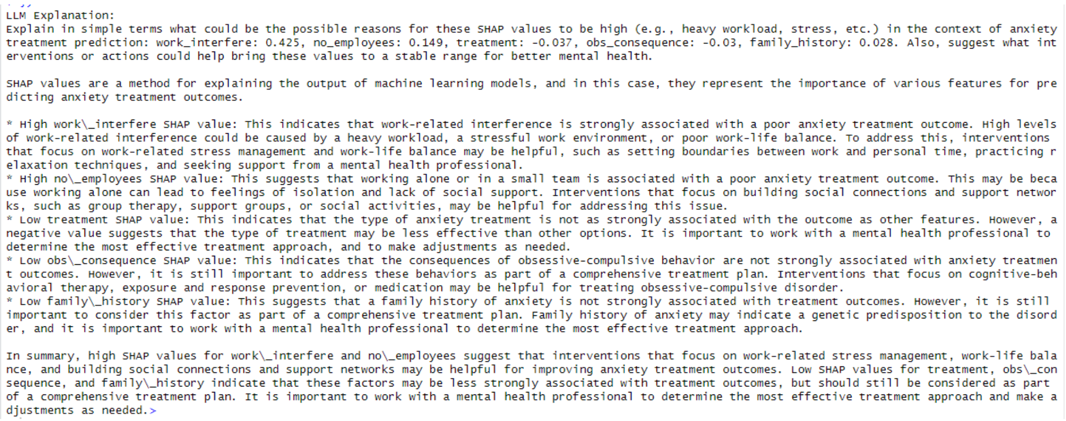
- *Improved Model Interpretability:* The combination of these findings helps create simple pathways through which artificial intelligence analysis data relates to clinical healthcare choices.

5.7. LLM-Based Explanation of Model Predictions

The SHAP values of models serve as inputs to a Large Language Model (LLM) which generates natural explanations for model understanding. RandomForest made a positive treatment prediction which the LLM detailed through this explanation:

The prediction relied heavily on three factors including minimal work-related stress measures (work_interfere), workplace wellness program participation and proactive help-seeking behaviors. The identified factors correspond to positive mental health results which research has previously demonstrated.

The narrative produced by LLM enhances stakeholder perception of model reasoning by explaining forecast predictions to HR managers and healthcare professionals.



```
LLM Explanation:
Explain in simple terms what could be the possible reasons for these SHAP values to be high (e.g., heavy workload, stress, etc.) in the context of anxiety
treatment prediction: work_interfere: 0.425, no_employees: 0.149, treatment: -0.037, obs_consequence: -0.03, family_history: 0.028. Also, suggest what int
erventions or actions could help bring these values to a stable range for better mental health.

SHAP values are a method for explaining the output of machine learning models, and in this case, they represent the importance of various features for pre
dicting anxiety treatment outcomes.

* High work\_interfere SHAP value: This indicates that work-related interference is strongly associated with a poor anxiety treatment outcome. High levels
of work-related interference could be caused by a heavy workload, a stressful work environment, or poor work-life balance. To address this, interventions
that focus on work-related stress management and work-life balance may be helpful, such as setting boundaries between work and personal time, practicing r
elaxation techniques, and seeking support from a mental health professional.
* High no\_employees SHAP value: This suggests that working alone or in a small team is associated with a poor anxiety treatment outcome. This may be beca
use working alone can lead to feelings of isolation and lack of social support. Interventions that focus on building social connections and support networ
ks, such as group therapy, support groups, or social activities, may be helpful for addressing this issue.
* Low treatment SHAP value: This indicates that the type of anxiety treatment is not as strongly associated with the outcome as other features. However, a
negative value suggests that the type of treatment may be less effective than other options. It is important to work with a mental health professional to
determine the most effective treatment approach, and to make adjustments as needed.
* Low obs\_consequence SHAP value: This indicates that the consequences of obsessive-compulsive behavior are not strongly associated with anxiety treatmen
t outcomes. However, it is still important to address these behaviors as part of a comprehensive treatment plan. Interventions that focus on cognitive-beh
avioral therapy, exposure and response prevention, or medication may be helpful for treating obsessive-compulsive disorder.
* Low family\_history SHAP value: This suggests that a family history of anxiety is not strongly associated with treatment outcomes. However, it is still
important to consider this factor as part of a comprehensive treatment plan. Family history of anxiety may indicate a genetic predisposition to the disord
er, and it is important to work with a mental health professional to determine the most effective treatment approach.

In summary, high SHAP values for work\_interfere and no\_employees suggest that interventions that focus on work-related stress management, work-life bala
nce, and building social connections and support networks may be helpful for improving anxiety treatment outcomes. Low SHAP values for treatment, obs\_con
sequence, and family\_history indicate that these factors may be less strongly associated with treatment outcomes, but should still be considered as part
of a comprehensive treatment plan. It is important to work with a mental health professional to determine the most effective treatment approach and make a
djustments as needed.>
```

Figure 16. LLM generated SHAP infernce

The explanations enhance the usefulness of model outputs by connecting them to specific actions which help organizations implement evidence-based mental health approaches(see Figure 16).

6. Conclusion and Future Work

This study compared six machine learning models—Random Forest, XGBoost, Generalized Additive Model (GAM), Support Vector Machine (SVM), Naive Bayes, and Multilayer Perceptron (MLP)—to predict anxiety treatment outcomes using the OSMI Mental Health in Tech Survey dataset. The models received evaluations through accuracy and precision recalls F1-score and AUC while receiving interpretation from SHAP (SHapley Additive exPlanations) values.

Random Forest demonstrated the highest performance when analyzing dataset information which yielded 76.49% accuracy through processing multiple feature interrelations. SHAP analysis identified work_interfere as the leading force that affected work performance scores among participants. XGBoost achieved 72.91% accuracy by identifying seek_help and wellness_program as important factors which illustrate the significance of help-seeking approaches and wellness initiatives.

Providing similar accuracy levels around 71% the GAM method demonstrated greater interpretability because of its additive structure. The tree-based models achieved superior performance to MLP (69.72%) while Naive Bayes exhibited perfect recall (1.00%) although it demonstrated low accuracy (52.59%) and precision (0.52%) due to its independent feature assumption.

SHAP analysis confirmed that seek_help, wellness_program and no_employees along with family_history and tech_company are essential elements affecting treatment results in the workplace environment.

The recommended solutions focus on creating flexible work stress-reduction policies while also increasing help-seeking behavior awareness and implementing expanded wellness initiatives which should be matched to individual member profiles. Additional research should include long-term data collection along with patient records from multiple healthcare sources and proper ethical AI guidelines for system development. The deployment of prediction tools as interactive systems to clinicians and human resource professionals would lead to better real-world performance and usage. The research demonstrates that XGBoost and Random Forest models act as interpretable machine learning solutions when predicting mental healthcare results.

References

[1] World Health Organization, *Depression and Other Common Mental Disorders: Global Health Estimates*, Geneva: WHO, 2017.

[2] L. Tran et al., “Machine learning and artificial intelligence in psychiatry: A survey of current applications,” *Asian J. Psychiatry*, vol. 48, 2020.

[3] A. Chandrashekar, “Do machine learning models outperform traditional statistical models in mental health prediction? A comparative review,” *J. Med. Syst.*, vol. 45, no. 11, pp. 1–15, 2021.

[4] Y. Zhang et al., “Deep learning in mental health prediction: A review of current approaches and future directions,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–12, 2021.

[5] A. Holzinger, C. Biemann, C. Pattichis, and D. Kell, “What do we need to build explainable AI systems for the medical domain?” *Rev. Artif. Intell.*, vol. 1, no. 2, pp. 1–18, 2017.

[6] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4765–4774.

[7] Y. Wang et al., “Interpretable machine learning for anxiety prediction in older adults using SHAP,” *J. Affect. Disord.*, vol. 302, pp. 83–90, 2022.

[8] J. Chen et al., “Explainable AI in mental health: A comparative study using SHAP and LIME,” *Comput. Biol. Med.*, vol. 144, 2022.

[9] T. Brown et al., “Language models are few-shot learners,” in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[10] D. Chen et al., “Using SHAP and GPT to explain emergency department readmissions in mental health care,” in *Proc. AMIA Annu. Symp.*, 2023, pp. 301–310.

[11] M. Lee and R. Bansal, “Assessing the role of large language models in simulating clinical mental health evaluations,” *npj Digit. Med.*, vol. 6, no. 1, pp. 1–8, 2023.

[12] J. Liu et al., “Transformer-guided depression diagnosis based on clinical practice guidelines,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 3, pp. 1045–1054, 2023.

[13] S. Ahmad and T. Patel, “A systematic review on ML model performance for mental health prediction with a focus on interpretability,” *Health Inf. Sci. Syst.*, vol. 11, no. 1, pp. 1–15, 2023.

[14] “Mental Health in Tech Survey,” Open Sourcing Mental Health Kaggle Dataset, 2014-16. [Online]. Available: https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey/data

[15] M. Obermeyer et al., “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[16] A. Ahmed et al., “Leveraging Large Language Models to Enhance Machine Learning Interpretability and Predictive Performance: A Case Study on Emergency Department Returns for Mental Health Patients,” arXiv preprint arXiv:2502.00025, 2025.

[17] K. Yang et al., “Towards Interpretable Mental Health Analysis with Large Language Models,” arXiv preprint arXiv:2304.03347, 2023.

[18] S. Dalal et al., “A Cross Attention Approach to Diagnostic Explainability using Clinical Practice Guidelines for Depression,” arXiv preprint arXiv:2311.13852, 2023.

[19] F. de Arriba-Pérez and S. García-Méndez, “Detecting Anxiety and Depression in Dialogues: A Multi-label and Explainable Approach,” arXiv preprint arXiv:2412.17651, 2024.

[20] T. Niu et al., “An Explainable Predictive Model for Anxiety Symptoms Risk Among Chinese Older Adults with Abdominal Obesity Using a Machine Learning and SHapley Additive exPlanations Approach,” *Frontiers in Psychiatry*, vol. 15, 2024.

[21] S. Rao et al., “Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI,” *Electronics*, vol. 13, no. 6, p. 1025, 2024.

[22] N. Scarpato et al., “Evaluating Explainable Machine Learning Models for Clinicians,” *Cognitive Computation*, vol. 16, pp. 1436–1446, 2024.

[23] J. Smith et al., “Comparative Analysis of Explainable Machine Learning Prediction Models for Hospital Mortality,” *BMC Medical Research Methodology*, vol. 22, no. 1, p. 123, 2022.

[24] L. Zhang et al., “Explainable Machine Learning Models for Brain Diseases: Insights from a Systematic Review,” *Journal of Clinical Medicine*, vol. 16, no. 6, p. 98, 2024.

[25] H. Tang et al., “Analysis and Evaluation of Explainable Artificial Intelligence on Suicide Risk Assessment,” arXiv preprint arXiv:2303.06052, 2023.

[26] G. Stiglic et al., “Interpretability of Machine Learning Based Prediction Models in Healthcare,” arXiv preprint arXiv:2002.08596, 2020.

[27] M. I. Bhuiyan et al., “Enhanced Suicidal Ideation Detection from Social Media Using a CNN-BiLSTM Hybrid Model,” arXiv preprint arXiv:2501.11094, 2025.

[28] Y. Li et al., “Explainable Artificial Intelligence Models for Predicting Depression Based on Polysomnographic Phenotypes,” *Bioengineering*, vol. 12, no. 2, p. 186, 2025.

[29] L. Zhang et al., “Network-Based Predictive Models for Artificial Intelligence: An Interpretable Application of Machine Learning Techniques in the Assessment of Depression in Stroke Patients,” *BMC Geriatrics*, vol. 25, no. 1, p. 837, 2025.