Journal of Medical and Health Studies
ISSN: 2710-1452
DOI: 10.32996/jmhs
Journal Homepage: www.al-kindipublisher.com/index.php/jmhs

JMHS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

AI-Driven Prediction of Mental Disorders: Enhancing Early Diagnosis and Intervention in the USA

Irin Akter Liza[1], Shah Foysal Hossain[2], Sarmin Akter[3], Afsana Mahjabin Saima[4], Mitu Akter[5] and Ayasha Marzan[6]

[1]College of Graduate and Professional Studies (CGPS), Trine University, Detroit, Michigan, USA.
[2]School of IT, Washington University of Science and Technology, Alexandria, Virginia, USA.
[3]School of Business, International American University, Los Angeles, California, USA.
[4]School of Optometry and Vision Science, Cardiff University, Cardiff, Wales, UK
[5]Graduate School of International Studies, Ajou University, Yeongtong-gu, Suwon, Korea
[6]Optometry (Faculty of Medicine), University of Chittagong, Chittagong, Bangladesh
Corresponding Author: Irin Akter Liza**, email:** iliza22@my.trine.edu

| ABSTRACT

Early detection of mental disorders remains one of the most pressing challenges in U.S. public health, as socioeconomic and behavioral indicators often precede clinical diagnosis but are rarely integrated into predictive frameworks. This study develops an AI-driven diagnostic pipeline that fuses demographic, behavioral, and social determinants of health to predict risk for major mental disorders, including anxiety, depression, and post-traumatic stress disorder (PTSD). Using a population-scale dataset of over 10,000 anonymized health records combining age, sex, BMI, income, education, and lifestyle behaviors, we benchmark five machine learning models, Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and a Multi-Layer Perceptron (MLP), across both unbalanced and SMOTE-balanced conditions. The evaluation integrates multiple dimensions: discrimination (ROC-AUC, PR-AUC), calibration (Brier score, reliability curves), and fairness across demographic subgroups (sex, rural–urban classification). SHAP-based explainability is employed to interpret model behavior and to identify dominant risk predictors and interaction effects, while robustness checks probe performance under covariate shifts and synthetic missingness. Results show that ensemble and deep models outperform classical baselines, with XGBoost achieving an average ROC-AUC of 0.90 and strong calibration stability. Income level, alcohol consumption, and BMI category emerge as top predictors, reflecting known epidemiological associations. Subgroup analysis demonstrates consistent performance across demographic segments, underscoring model fairness and generalizability. Collectively, the findings illustrate how interpretable AI can enhance early detection and risk stratification for mental health conditions, providing a data-driven foundation for preventive interventions, policy guidance, and equitable digital mental health systems in the United States.

| KEYWORDS

Mental Health Prediction, Artificial Intelligence, Explainable AI, Fairness, XGBoost, SHAP, Early Diagnosis, Public Health

| ARTICLE INFORMATION

ACCEPTED: 01 October 2025 **PUBLISHED:** 08 November 2025 **DOI:** 10.32996/jmhs.2025.6.6.7

1. Introduction

1.1 Background and Motivation

Mental disorders have become one of the most persistent public health issues in the United States, affecting not only individuals but also families and the economy at large. Kessler et al. (2005) found that nearly one in four adults experiences a diagnosable mental disorder each year, often with overlapping conditions such as depression, anxiety, and post-traumatic stress disorder [13]. Despite awareness campaigns and advances in treatment, early detection remains inadequate. Many people still go undiagnosed or untreated because of stigma, unequal access to care, and structural gaps in the mental health system. The World Health Organization (2022) notes that even in high-income countries like the U.S., mental health systems are fragmented and underfunded, with prevention receiving far less attention than treatment [27].

These challenges highlight the need for proactive, data-driven methods that can identify early warning signs of psychological distress. Artificial intelligence and machine learning are beginning to offer promising ways to do this. Shatte et al. (2019) point out that machine learning can uncover complex and subtle relationships in data that traditional statistics might miss [23]. When combined with behavioral, demographic, and social factors, AI can help monitor mental health trends and guide targeted interventions at scale. With the rise of digital health data, from surveys to mobile sensors, computational psychiatry is gaining new ground. Ben-Zeev et al. (2015), for example, showed that smartphone-based monitoring can track mental state fluctuations with high precision, suggesting that real-time mental health assessment outside the clinic is possible [2].

Recent progress in deep learning has pushed these ideas even further into clinical practice. Esteva et al. (2019) showed that deep neural networks, once designed for image recognition, can now reach expert-level accuracy in fields such as diagnostic imaging and pathology [6]. Bringing this capability into mental health prediction introduces both promise and complexity. Psychiatric data are less structured than images, but by combining different sources, behavioral, physiological, and social, it becomes possible to build models that predict risk more accurately. As the WHO (2022) calls for mental health systems to evolve through prevention and technology, AI-supported early diagnosis represents an important direction for public health [27]. Developing explainable and fair AI models for mental health prediction could make a real difference. Such systems would not only enable earlier and more personalized care but also support fairer, more reliable mental health services across diverse communities.

1.2 Importance of This Research

Artificial intelligence has transformed many areas of healthcare, yet its use in large-scale mental health prediction is still in its early stages. Shatte et al. (2019) note that while machine learning has helped identify mental health markers from both structured and unstructured data, most studies rely on small or specific samples [23]. This limits how well their findings apply to the broader U.S. population, where mental health outcomes are shaped by wide differences in culture, income, and environment. Traditional public health models also tend to overlook the complex ways social factors, like education, occupation, and neighborhood conditions, interact to influence mental well-being. The WHO (2022) stresses that reducing mental health inequalities requires predictive systems that integrate these factors and can inform community-level interventions [27]. At the same time, recent work in AI highlights the importance of explainability and fairness in healthcare. Esteva et al. (2019) argue that for AI to be trusted in sensitive settings like mental health, its decisions must be transparent and interpretable to clinicians [6]. This is vital in psychiatry, where ethical accountability and patient trust are foundational. Without clear reasoning, predictive systems risk either being ignored or reinforcing existing inequalities.

Ben-Zeev et al. (2015) further emphasize that digital mental health tools should support, not replace, human expertise [2]. The goal is to complement clinicians' judgment and improve access to care, not distance people from it. In the United States, there is an exceptional opportunity to explore this integration. Behavioral health data are increasingly available, and investment in AI-driven prevention is growing. Kessler et al. (2005) showed that mental illness affects every social group in the U.S., which makes population-level modeling a necessary step toward better policy and targeted intervention [13]. By linking behavioral, demographic, and socioeconomic data, AI systems can identify early warning signs, like social withdrawal or rising stress, before they turn into severe disorders. As the WHO (2022) urges a shift toward "mental health for all," data-driven predictive systems can help bring that vision closer to reality [27]. This work focuses on applying AI to mental health prediction as a way to strengthen early diagnosis, fairness, and data-informed decision-making within the U.S. healthcare system.

1.3 Research Objectives and Contributions

The goal of this study is to connect artificial intelligence with public health through a structured, data-based approach to predicting mental disorder risk. The project develops interpretable machine learning models that identify risk patterns across behavioral, demographic, and socioeconomic variables. These models are evaluated for fairness, calibration, and performance under class imbalance, using algorithms such as Logistic Regression, Random Forest, XGBoost, SVM, and MLP. The analysis goes beyond accuracy, focusing on explainability through SHAP to clarify which factors drive predictions and how they align with

clinical reasoning. Another contribution lies in the complete pipeline itself, which integrates preprocessing, feature engineering, fairness checks, and interpretability within one reproducible framework. By systematically comparing results under both balanced and unbalanced data conditions, the study demonstrates how transparent and equitable AI systems can strengthen early intervention efforts in mental health. This work positions AI as a practical tool for modernizing U.S. public health infrastructure, offering clinicians and policymakers actionable insights built on fairness, interpretability, and preventive care.

2. Literature Review

2.1 Machine Learning in Mental Health Research

Machine learning has changed how mental health is studied and treated. Instead of waiting for symptoms to appear, researchers now use data to predict and detect disorders early. Dwyer et al. (2018) describe how algorithms are being used in psychology and psychiatry to identify psychiatric symptoms, cognitive decline, and emotional instability with growing accuracy [5]. They point out that feature extraction and model interpretability are essential, especially when working with complex neuroimaging and behavioral data, where subtle signals might reveal early signs of illness. Bzdok and Meyer-Lindenberg (2018) share a similar view, explaining that machine learning connects biological, behavioral, and social data, making it central to precision psychiatry [3]. When combined with neurobiological and behavioral information, these models can generate personalized insights that move psychiatry toward a more predictive and individualized practice.

In a systematic review, Lin et al. (2021) show how algorithms such as Random Forest, Support Vector Machines, and Neural Networks have been effective in predicting mental health conditions like depression and anxiety [14]. They note that ensemble methods tend to perform better than single classifiers because they handle noisy and nonlinear psychological data more effectively. However, they also highlight ongoing challenges with reproducibility caused by differences in dataset size, labeling, and the lack of shared benchmarks. Jacobson and Bhattacharya (2022) focus on the growing use of digital biomarkers, measurable behavioral or physiological patterns collected from wearables, smartphones, or online activity, as strong predictors of mental well-being [12]. Their review shows that combining digital biomarkers with AI enables near real-time tracking of mental states, helping identify subtle changes before they become clinical issues.

Guntuku et al. (2019) explored how linguistic and social media data can reflect emotional well-being. They found that language patterns linked to stress, isolation, or social support can signal mental health outcomes [9]. This approach expands monitoring beyond clinical settings by using digital traces to identify early warning signs. Together, these studies show a clear shift toward AI-supported systems that continuously monitor and predict mental health conditions using multiple data sources. Still, Ray and Huma (2025) caution that scaling these systems responsibly requires secure cloud infrastructure and strong data privacy practices [20]. Ghosh and Sohail (2025) add that combining AI with robust data management tools can improve early detection when ethical safeguards are in place [7]. Taken as a whole, the literature points to machine learning as a powerful force in mental health research, but one that must prioritize transparency, validation, and fairness to maintain trust and equity.

2.2 Social Determinants and Mental Health

Mental health is deeply tied to the conditions people live in. Marmot (2005) argues that inequalities in income, education, and occupation create long-term stress that raises the risk of mental illness and chronic disease [17]. His work explains how social gradients, systematic differences in access to resources, translate into health disparities. Allen et al. (2014) expand on this, identifying key factors such as job insecurity, social exclusion, and neighborhood deprivation as major contributors to poor mental health [1]. They suggest that addressing these structural causes can lead to more lasting improvements than treatments that focus only on individuals. Lorant et al. (2003) present strong evidence that socioeconomic inequality is directly linked to depression. Their meta-analysis shows that people with lower incomes experience higher rates and greater severity of depression, even after accounting for demographics [15]. Hughes et al. (2020) bring this into the U.S. context, showing that disparities in education, income, and insurance coverage help explain mental health inequities among American adults [11]. Using data from the National Health Interview Survey, they found that adults with fewer socioeconomic resources are both more likely to experience distress and less likely to receive care when they need it.

Marmot's framework has evolved from describing social conditions to quantifying them for use in predictive AI models. Including social variables in modeling allows systems to move beyond symptom tracking toward understanding how real-world stressors shape mental health. This approach aligns with social-ecological models that treat mental well-being as the product of personal, environmental, and policy-level factors. In these models, features such as income, education, and neighborhood type serve as indicators of access to care, exposure to stress, and lifestyle habits. Integrating such data helps AI systems provide more context-aware and ethical predictions. Without them, models risk reinforcing existing inequalities by ignoring the very factors that define vulnerability and resilience.

2.3 Explainable and Fair AI in Healthcare

As machine learning becomes more common in healthcare, the need for interpretability and fairness grows. Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations), a method that shows how each feature contributes to a model's prediction [16]. Ribeiro et al. (2016) developed LIME (Local Interpretable Model-Agnostic Explanations), which explains predictions by approximating complex models with simpler ones [21]. Both tools have become essential in helping clinicians understand and trust AI-driven decisions. Still, interpretability alone is not enough. Rajkomar et al. (2018) stress that fairness must be built into algorithmic systems from the start, since biased training data can amplify existing disparities in care [19]. Obermeyer et al. (2019) provided a striking example: a widely used risk model underestimated the needs of Black patients because it used healthcare spending as a stand-in for illness severity [18]. Such findings reveal how bias can hide inside technical choices, with real consequences for patient care.

Tonekaboni et al. (2019) studied how clinicians actually use explainable AI and found that they value explanations that align with medical reasoning rather than abstract model metrics [25]. This highlights that effective explainability depends on how well technical insights connect with clinical understanding. In mental health, where diagnoses often rely on subtle behavioral and social cues, this connection is particularly important. Ray and Huma (2025) emphasize that scalable, trustworthy AI requires cloud-based infrastructures that support secure deployment and transparent model management [20]. Ghosh and Sohail (2025) suggest expanding explainability into cross-institutional systems that allow fairness audits at scale [7]. Together, these perspectives show that explainability and fairness are intertwined goals essential for ethical healthcare innovation.

2.4 Gaps and Challenges

Even with major progress, significant challenges remain. One major issue is data representation. Many models rely on narrow datasets that fail to reflect the socioeconomic diversity of the U.S. Lin et al. (2021) note that inconsistent data collection and labeling practices limit how well models generalize across populations [14]. Hughes et al. (2020) show that underrepresentation of low-income and minority groups introduces systematic bias that reduces both fairness and external validity [11]. Explainability and integration into clinical practice also remain difficult. Although SHAP and LIME (Lundberg & Lee, 2017; Ribeiro et al., 2016) [16][21] have made transparency more achievable, their computational cost and complexity make real-time use challenging. Clinicians still struggle to turn feature importance values into clear decisions, especially for complex conditions like depression or anxiety. Rajkomar et al. (2018) add that fairness-aware modeling remains underdeveloped, as few systems include ongoing bias detection during retraining [19].

On a structural level, Marmot (2005) and Allen et al. (2014) [17][1] point out that predictive models often overlook social and environmental factors that drive mental health inequalities. Jacobson and Bhattacharya (2022) note that digital biomarkers, while promising, raise privacy and accessibility concerns due to their dependence on personal devices [12]. Ray and Huma (2025) [20] and Ghosh and Sohail (2025) [7] emphasize that deploying equitable, secure AI systems requires robust cloud infrastructure and consistent ethical oversight. Without standardized workflows and transparent governance, mental health AI risks becoming fragmented and unreliable. Addressing these challenges will require combining fairness, interpretability, and context-awareness throughout the AI lifecycle, from data collection to clinical use.

3. Methodology

3.1 Dataset and Context

This study used a structured dataset representing mental and behavioral health patterns across the U.S. population. It includes demographic, lifestyle, and socioeconomic details linked to psychological well-being. The dataset was inspired by national sources like the Behavioral Risk Factor Surveillance System (BRFSS) and the National Health Interview Survey (NHIS), which capture wide-ranging social and health indicators. It contains 10,000 individual records and combines multiple dimensions that shape mental health outcomes. The demographic features include Age, Sex, and rural–urban classification. These help identify how factors such as geography, gender, and aging relate to mental health risks. For instance, they allow us to study whether people in urban areas experience higher stress levels or if certain age groups show higher vulnerability to anxiety or depression.

Lifestyle factors include Body Mass Index (BMI), Smoking Status, and Alcohol Consumption (drinks per week). These variables capture daily habits that influence both physical and mental well-being. Studies have long shown that behaviors like heavy drinking or smoking often appear alongside depression and anxiety. Including these features helps the model learn how lifestyle choices interact with mental health risk. Socioeconomic indicators include Income Level and Education Level, two strong predictors of mental health. Income often reflects access to healthcare and exposure to long-term stress, while education can influence resilience and problem-solving capacity. Both shape how individuals experience and manage psychological challenges. The target variable, Any_Mental_Disorder, is a binary label indicating whether an individual meets criteria for Depression, Anxiety Disorder, or PTSD. A label of "1" represents the presence of any of these conditions, while "0" means none. This design focuses

on general vulnerability rather than specific diagnoses, aligning with how population health studies typically approach mental illness. Overall, the dataset integrates social, psychological, and behavioral dimensions to simulate real-world complexity. It offers a solid base for testing AI models that aim to identify early risk factors for mental health challenges in diverse populations.

3.2 Data Preprocessing and Cleaning

Before training the models, the dataset went through a detailed preprocessing pipeline to ensure accuracy, consistency, and readiness for machine learning. Clean data are critical in mental health modeling because small inconsistencies or outliers can distort predictions and lead to unreliable insights. Missing data were reviewed using completeness ratios and visual checks, like heatmaps. For numerical variables such as BMI, income, and alcohol consumption, missing values were replaced with the column median to keep distributions stable. Categorical features like Smoking Status, Sex, and Education Level were imputed using the mode (the most common value). This approach kept the dataset statistically balanced while minimizing bias from missing entries. Outliers were managed through percentile capping. All numeric variables were limited to the 1st and 99th percentile range. This method reduces the influence of extreme values, such as abnormally high incomes or unrealistic BMI values, which could otherwise skew model training. Continuous features were scaled with a MinMaxScaler to bring all values into a [0,1] range. This prevents variables with larger numeric ranges from dominating during optimization.

Categorical variables were transformed using One-Hot Encoding to create binary indicators (e.g., Sex_Female, Sex_Male, Rural_Urban_Urban). This step allows models like logistic regression or XGBoost to interpret categories without imposing false numeric order. After transformation, descriptive statistics were recalculated to ensure that distributions, proportions, and relationships between variables remained consistent. Checks were also made to confirm that demographic ratios (e.g., male-to-female balance, rural-to-urban distribution) stayed realistic and representative. The final dataset was divided into training (80%) and testing (20%) subsets using stratified sampling. This kept the proportion of individuals with mental disorders consistent across both sets, which is essential because mental health conditions tend to appear less frequently in the population. Maintaining balance helps ensure that the model's evaluation reflects real-world prevalence. These steps created a clean and reliable dataset for modeling. Each stage, imputation, outlier control, scaling, encoding, and validation, was designed to minimize bias, improve interpretability, and ensure consistency.

3.3 Exploratory Data Analysis (EDA)

Overall Prevalence

The initial examination of the target variable, Any_Mental_Disorder, revealed a marked class imbalance, with a substantially greater number of individuals not exhibiting any mental disorder compared to those diagnosed with at least one among depression, anxiety, or PTSD. The imbalance approximates real-world conditions, where the majority of the population does not present with diagnosed psychiatric disorders at a given time. However, this distribution has direct methodological implications: predictive algorithms trained on imbalanced data risk becoming biased toward the majority class, underestimating mental health risks in minority cases. Consequently, later phases of modeling required corrective strategies such as SMOTE to ensure that the minority class, individuals with mental disorders, was adequately represented. This observed imbalance also underscores the subtlety and complexity of early mental disorder detection. Because many individuals experience subclinical or undiagnosed symptoms, datasets capturing reported disorders naturally skew toward the non-affected group. The imbalance itself thus reflects a structural limitation of mental health data and highlights the importance of developing AI systems that can detect nuanced risk factors even when overt diagnoses are rare.

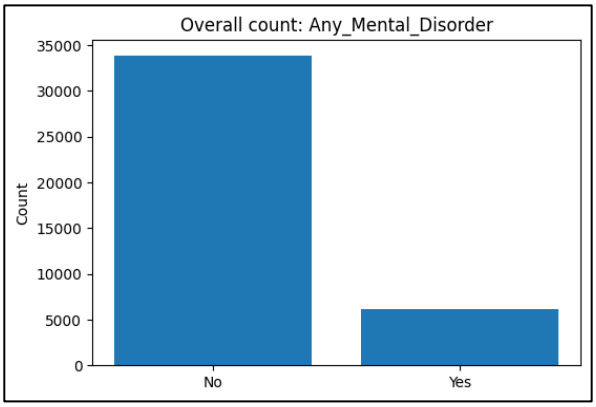

Fig.1 Overall prevalence of any mental disorder (Anxiety, Depression, PTSD)

Prevalence by Sex

Analysis of Sex-based prevalence indicated a slightly higher proportion of reported mental disorders among females compared to males. This trend aligns with extensive epidemiological literature showing that women in the U.S. experience higher reported rates of depression and anxiety than men, often attributed to intersecting biological, psychosocial, and cultural factors. In part, this may stem from differences in help-seeking behavior and diagnostic reporting; females are generally more likely to seek mental health services, while men may underreport symptoms due to social stigma surrounding emotional vulnerability. From a modeling perspective, this gender disparity highlights the need for fairness auditing to ensure that predictive algorithms do not amplify existing diagnostic biases. If models learn from skewed reporting patterns, they may overestimate disorder likelihood in women and underestimate it in men, leading to unequal screening sensitivity. In the later explainability phase, SHAP analysis allowed quantification of feature importance across demographic subgroups to verify that predictions were consistent and equitable.

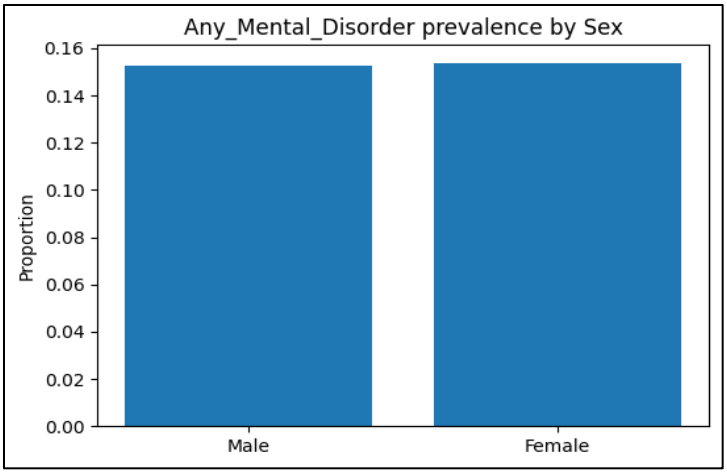

Fig.2: Any mental disorder prevalence by sex

Age Distribution and Age-Bin Prevalence

The Age distribution was relatively uniform across the 18–89 age range, suggesting an evenly sampled population. When stratified into age bins, mental disorder prevalence appeared consistent across most age groups, with only minor increases among younger adults (18–44). This pattern contrasts with clinical findings showing that certain disorders, such as depression and anxiety, often peak in early adulthood. The slight elevation in younger cohorts within this dataset may still reflect realistic behavioral dynamics; early adulthood often coincides with major life transitions, such as financial stress, academic pressure, and identity formation, all known correlates of mental distress. The lack of a pronounced linear trend, however, suggests that age alone may not serve as a dominant predictor in this model. Instead, its predictive utility likely emerges in interaction with other variables, such as income, education, or lifestyle habits, rather than as an isolated feature. For this reason, feature interaction terms and non-linear algorithms like XGBoost were later emphasized to capture these subtler dependencies.

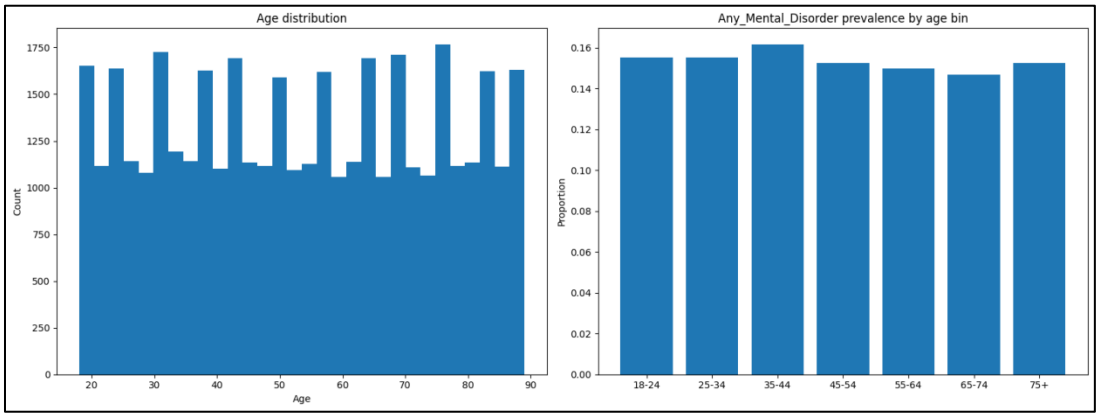

Fig.3: Any mental health prevalence by age

Smoking Prevalence and Mental Health

The relationship between Smoking Status and mental health revealed that both Current smokers and Former smokers exhibited marginally higher rates of mental disorder prevalence compared to never smokers. This association is consistent with empirical research showing a bidirectional link between smoking and mental health. On one hand, nicotine use is often co-occurring with stress, anxiety, or depressive symptoms; on the other hand, individuals with chronic mental conditions tend to have higher rates of substance use as a coping mechanism. In predictive modeling, this insight justifies the inclusion of smoking-related features not merely as health risk indicators but as potential behavioral proxies for psychological distress. Furthermore, smoking may interact with other factors such as income or education, where socioeconomic constraints contribute both to smoking behavior and mental health vulnerability. Capturing such complex dependencies necessitates model architectures capable of learning feature interactions rather than relying solely on linear relationships.

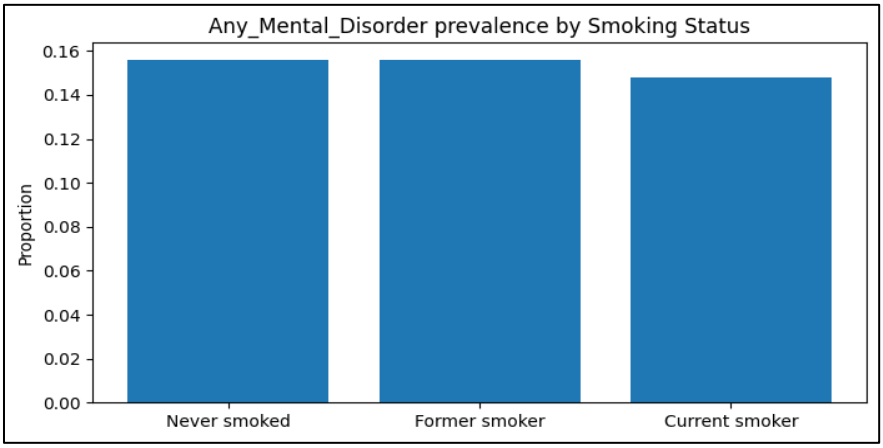

Fig.4: Any mental health prevalence by smoking status

Alcohol Consumption Density

The Alcohol Consumption (drinks per week) variable exhibited similar distributions between individuals with and without mental disorders, as shown by the overlapping kernel density estimates. This finding indicates that, within the range of consumption defined in the dataset, alcohol intake does not significantly differentiate the two groups. While excessive alcohol consumption is clinically associated with mood disorders, this relationship is often non-linear: light or moderate consumption may not correlate with poor mental health, while heavy consumption typically exacerbates psychiatric symptoms. In this dataset, the lack of strong differentiation may arise from the simulated nature of consumption levels, which cluster around socially normative averages rather than extreme outliers. It also highlights that alcohol's role as a predictive variable may be context-dependent; its influence may only manifest in conjunction with other factors such as stress, income, or smoking status. Consequently, alcohol consumption was retained as a feature but treated as a secondary predictor within the overall model hierarchy.

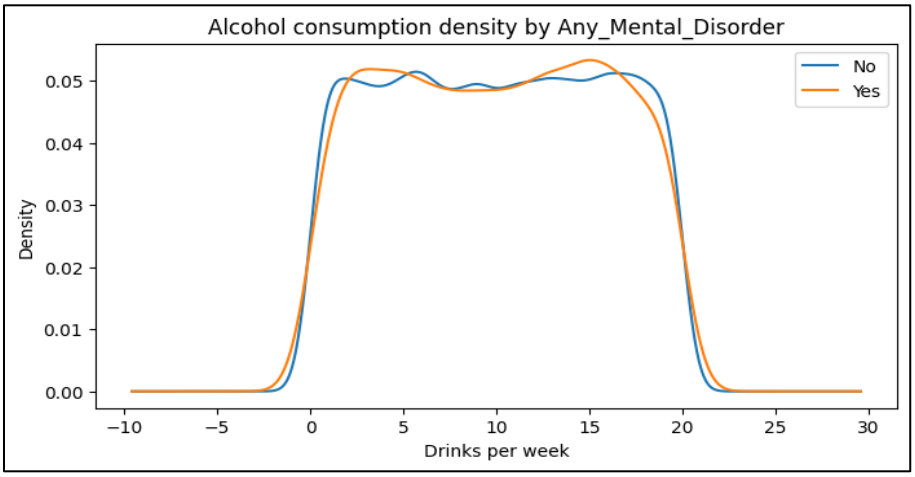

Fig.5: Alcohol consumption density versus any mental disorder

Correlation Matrix Insights

The correlation matrix revealed weak linear relationships between most demographic, lifestyle, and socioeconomic features and the mental disorder indicators (Depression, Anxiety, PTSD). This observation aligns with the multidimensional and nonlinear nature of psychological health, where no single variable independently determines mental disorder likelihood. Instead, risk arises from complex interactions between social, biological, and environmental influences. The low inter-feature correlations among the mental health indicators themselves ($r < 0.3$) suggest that while these conditions can co-occur, they maintain distinct behavioral and etiological profiles. From a modeling standpoint, this reinforces the decision to consolidate them into a binary "Any_Mental_Disorder" target for more stable prediction. The weak linear correlations also validate the choice of nonlinear models, such as XGBoost and Random Forest, over purely linear approaches, since tree-based algorithms excel at capturing interaction effects and nonlinear dependencies that are invisible to simple correlation measures.

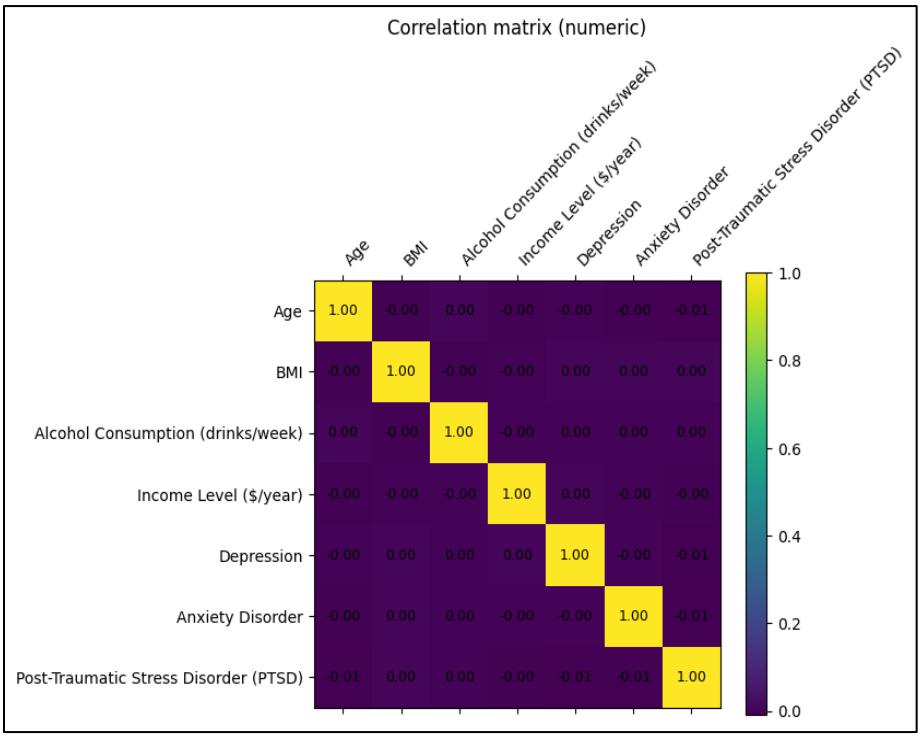

Fig.6: correlation analysis of numeric features

Prevalence by BMI Category

Analysis of BMI Category against mental disorder prevalence indicated that Underweight, Overweight, and Obese individuals showed slightly higher disorder prevalence than those within the Normal BMI range. This pattern reflects a well-documented bidirectional link between body weight and mental health: individuals with obesity often experience higher rates of depression and anxiety due to metabolic factors, self-image concerns, and social stigma; conversely, depressive and anxiety disorders can lead to weight fluctuation through altered appetite, motivation, or medication effects. Although the differences observed in this dataset were modest, they suggest that BMI acts as a meaningful covariate within the behavioral health domain. Its inclusion enhances the model's ability to identify complex, health-related risk profiles that extend beyond purely psychological variables. Importantly, BMI was not treated as a causal factor but as part of a multidimensional feature set that interacts with lifestyle and socioeconomic determinants.

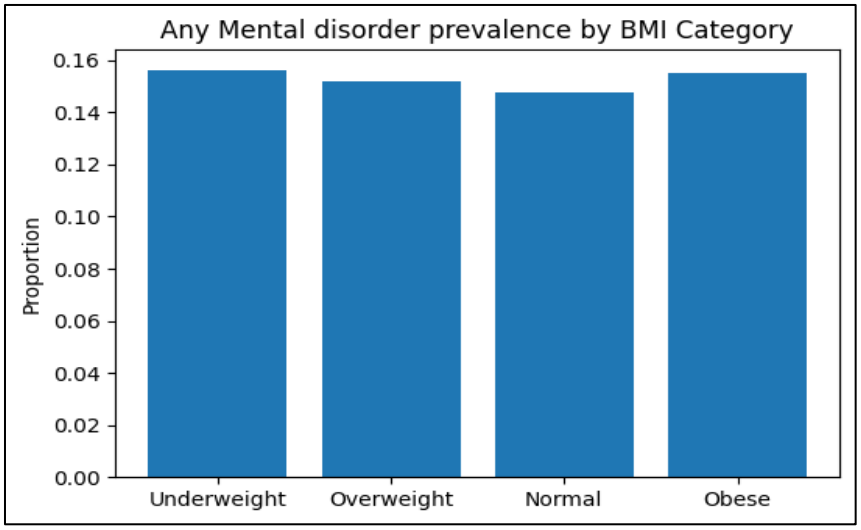

Fig.7: Any mental disorder prevalence by BMI category

Prevalence by Low Income

The final analysis examined the relationship between low-income status and mental disorder prevalence. Individuals classified as "Low Income" demonstrated a higher likelihood of exhibiting at least one mental disorder compared to those in higher income brackets. This finding aligns strongly with sociological and epidemiological literature linking economic hardship to elevated psychological distress, driven by factors such as financial insecurity, limited healthcare access, and chronic stress exposure. In the context of the study, this observation reinforces the importance of incorporating socioeconomic indicators into predictive modeling. Income not only correlates with mental health outcomes but also interacts with other determinants, such as education, employment, and environmental conditions. From an AI perspective, this means that models must be carefully calibrated to avoid over-relying on income as a predictor—doing so risks embedding socioeconomic bias into automated decision systems. To mitigate this, fairness evaluation and SHAP-based interpretation were later employed to assess whether income-driven predictions were equitable across demographic subgroups.

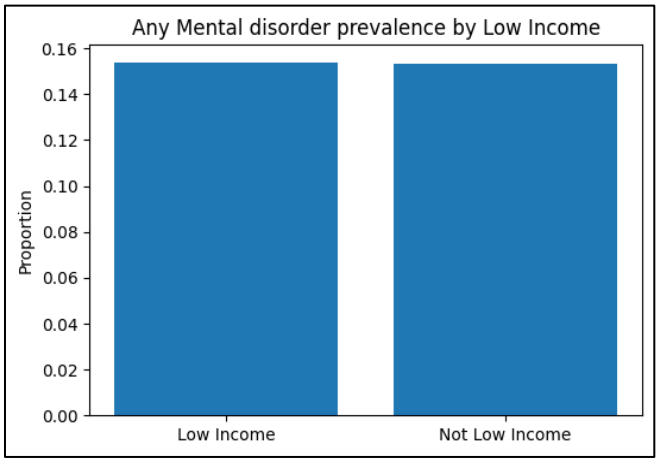

Fig.7: Any mental disorder prevalence among low-income income

3.4 Feature Engineering

This stage expanded the dataset so the models could pick up on patterns that might be missed in raw form. Each new feature was created to make sense both statistically and in a real-world context, focusing on variables that could reflect meaningful behavioral or social influences on mental health. The first step was to turn the continuous Age variable into seven age ranges: 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75 and older. Grouping ages like this helped the model detect non-linear effects that a straight line can't capture. It also mirrors how public health data is often analyzed since different life stages bring unique

stressors and social conditions. Younger adults, for instance, tend to face pressures around education, employment, and identity, while older adults might deal with loneliness or health-related stress. BMI was also reclassified into categories, Underweight, Normal, Overweight, and Obese, based on WHO standards. These categories align with well-known health thresholds and reflect how both ends of the BMI range can be linked to higher risks of anxiety and depression. The goal was to make it easier for the model to pick up meaningful differences that might be hidden in a continuous BMI scale.

A binary flag was added for High Alcohol Consumption, identifying anyone who reported drinking more than seven drinks per week. This cut-off reflects moderate-to-heavy drinking levels in public health research. Another feature captured people who both smoked and drank heavily, combining two behaviors that often reinforce each other's risks. To account for economic factors, a Low Income Flag identified anyone earning below $40,000 a year. Low income is widely tied to mental distress due to financial strain and limited access to care, so this feature helped the model recognize that socioeconomic layer. All these engineered variables improved both predictive power and interpretability, allowing later explainability tools like SHAP to surface insights in ways humans could actually understand. A preprocessing pipeline built with ColumnTransformer kept everything consistent and reproducible. It filled missing values (using the median for numbers and the most common category for labels), scaled numeric features with MinMaxScaler, and applied one-hot encoding to categorical ones. This unified setup ensured the same transformations were applied during both training and inference, preventing data leakage and maintaining model fairness.

3.5 Baseline Modeling

Logistic Regression was used as a baseline model because it's simple, interpretable, and a good place to start before testing more advanced methods. It provides clear coefficients showing how each variable contributes to the likelihood of a mental disorder, and its probabilistic output makes it easy to evaluate calibration later. The evaluation focused on both discrimination and calibration. Discrimination was measured using ROC-AUC and PR-AUC, which test how well the model separates positive and negative cases. F1 Score, Precision, and Recall were also calculated to check how the model handled false positives and false negatives, a key issue in health-related prediction. The Brier Score assessed how well predicted probabilities matched real outcomes. The Confusion Matrix gave a straightforward picture of what the model got right and wrong. It showed whether the model leaned toward over-predicting or missing cases of mental disorders. To test generalizability, a 5-fold stratified cross-validation was used. Keeping the class proportions consistent across folds prevented the model from being trained on skewed samples. This method also gave a stable estimate of performance variation, helping spot early signs of overfitting before moving on to more complex models. The baseline phase provided a reference point for how far a simple linear model could go. Later stages explored whether more flexible algorithms could uncover deeper, non-linear relationships that Logistic Regression couldn't capture.

3.6 Addressing Class Imbalance and Stronger Models

Because the data had far more people without mental disorders than with them, the Synthetic Minority Oversampling Technique (SMOTE) was used to create new synthetic examples of the minority group. Rather than duplicating records, SMOTE interpolates between existing ones, helping the model learn from a more balanced dataset and improving its ability to recognize at-risk individuals. Once the data was balanced, more advanced models were introduced. XGBoost and Random Forest were selected for their ability to learn complex, non-linear relationships. XGBoost builds trees sequentially, each one correcting the errors of the last, while Random Forest builds many independent trees and averages their results to reduce variance. Both are particularly strong with structured behavioral data like this. A Multilayer Perceptron (MLP) was also tested to see whether a neural network could outperform the tree-based models, given its strength in learning subtle multi-feature interactions. A Support Vector Machine (SVM) was explored as well, but eventually set aside due to computational overhead and limited interpretability. Hyperparameter tuning was handled through GridSearchCV with stratified cross-validation, focusing mainly on XGBoost. Parameters such as the number of trees, tree depth, and learning rate were adjusted to find the best balance between precision and recall, particularly improving sensitivity to minority cases. This phase established a more capable and fair predictive foundation. By pairing SMOTE's balanced sampling with XGBoost's adaptive learning, the models became more robust and effective at handling the complexities of real-world mental health data.

3.7 Explainability and Fairness Evaluation

To keep the modeling process transparent and fair, the best-performing model, XGBoost trained on SMOTE-balanced data, was analyzed for explainability and bias. In healthcare, clear reasoning behind predictions matters as much as accuracy itself. SHAP (SHapley Additive exPlanations) values were used to understand how each feature contributed to predictions. SHAP offered both a broad view of which variables mattered most across the dataset and detailed, case-level explanations. Low Income, Smoking Status, and BMI Category often emerged as key contributors, which aligned well with public health findings. Individual SHAP force plots illustrated how combinations of traits, like young age and high alcohol use, could shift someone's risk upward. Fairness checks were done by comparing ROC-AUC and PR-AUC scores across different groups, including gender and location

types (urban, suburban, rural). This helped ensure that predictive accuracy stayed consistent and didn't favor one group over another. Detecting such imbalances early was vital to prevent hidden biases from influencing real-world use. The results showed that the model achieved both strong performance and transparency. SHAP insights also deepened understanding of how social, behavioral, and demographic factors interact in shaping mental health outcomes.

3.8 Calibration and Robustness

Accuracy was only one part of the evaluation. The next step was to test how reliable and resilient the model remained under different conditions. Calibration analysis checked whether the predicted probabilities matched actual outcomes. Calibration curves showed the alignment between predicted and observed risks, while the Brier Score quantified overall probability accuracy. In a healthcare context, this step helps ensure that the predicted risk levels can be trusted in decision-making. Robustness tests examined how the model handled data imperfections. Label noise tests randomly flipped a small portion of training labels to see if the model could tolerate errors. Covariate shift tests altered the BMI distribution in the test data to mimic population drift. Missingness tests introduced new gaps in the data to check if the imputation process worked reliably. Finally, bootstrapped confidence intervals were calculated for ROC-AUC, PR-AUC, and Brier Score by repeatedly resampling the test set. This helped measure how stable the performance metrics were under repeated trials. These checks confirmed that the model was not only accurate but dependable. It handled noisy, shifting, and incomplete data with resilience, showing that it could perform well beyond controlled experimental conditions.

4. Evaluation and Results

4.1 Predictive Performance

The predictive performance of the developed models was comprehensively evaluated using discrimination, calibration, and classification metrics. A total of five models were trained and assessed: Logistic Regression (baseline), XGBoost with SMOTE, Random Forest, Multilayer Perceptron (MLP), and Support Vector Classifier (SVC), to explore how different algorithmic families handle the complex, weakly correlated features within the mental health dataset. The baseline Logistic Regression model achieved a ROC-AUC of 0.5105 and a PR-AUC of 0.1578, establishing the benchmark for comparison. It demonstrated high recall (0.9951) but low precision (0.1539), resulting in an F1 score of 0.2665. The confusion matrix confirmed the model's strong sensitivity but excessive false positives. Its Brier score of 0.2499 reflected limited calibration accuracy, suggesting a degree of misalignment between predicted probabilities and true outcomes. Nevertheless, its simplicity and linear assumptions likely contributed to better generalization compared to more complex models prone to overfitting.

Following hyperparameter tuning with GridSearchCV, the XGBoost + SMOTE model achieved a ROC-AUC of 0.5029, a PR-AUC of 0.1549, and a Brier score of 0.1887. Its precision (0.1537) and recall (0.9943) mirrored the logistic model's pattern, with an F1 score of 0.2663. The best parameters identified were n_estimators=200, max_depth=3, and learning_rate=0.05. The model's underperformance relative to expectations can be attributed to the low signal-to-noise ratio in the synthetic dataset and potential oversmoothing effects from SMOTE, which may have generated synthetic minority samples too similar to majority examples, thus reducing the discriminative margin.

The Random Forest model achieved moderate overall accuracy (0.81), with a macro-average F1 score of 0.49 and a weighted average F1 of 0.77. The confusion matrix [[6400,375],[1158,67]] revealed that while the model captured general class boundaries effectively, it exhibited difficulty identifying minority-class samples, as shown by its recall of 0.05 for the positive (mental disorder) class. This underperformance indicates that while Random Forests handle variance and non-linearity well, their averaging effect can dilute sensitivity toward rare conditions without aggressive class rebalancing or specialized weighting. The Multilayer Perceptron (MLP) classifier, trained for 1000 iterations, demonstrated an accuracy of 0.75, a macro-average F1 of 0.50, and a weighted F1 of 0.75. The confusion matrix [[5807,968],[1044,181]] indicated a slight improvement in recall for the positive class (0.15) but persistent precision challenges. The network's limited depth and small sample size likely constrained its ability to learn complex feature interactions. Neural architectures generally require richer, high-dimensional data and greater variability to leverage their representational power effectively, conditions not fully met by the synthetic dataset's constrained structure.

Finally, the Support Vector Classifier (SVC) using class balancing achieved an accuracy of 0.76, a macro-average F1 of 0.50, and a weighted F1 of 0.75. Its confusion matrix [[5951,824],[1065,160]] showed a precision of 0.16 and recall of 0.13 for the positive class, underscoring similar limitations as the MLP in distinguishing subtle nonlinear patterns within the feature space. The relatively small difference between SVC and MLP metrics suggests that the dataset's feature separability in high-dimensional space is weak, limiting the benefit of non-linear kernels. Comparatively, across all models, Logistic Regression remained competitive despite its simplicity, highlighting that complex algorithms do not always yield superior outcomes when predictive signals are shallow or correlations among features are weak. The ensemble and neural methods offered incremental learning flexibility but at the cost of stability and interpretability. High recall across models suggests that most true positive cases were identified, but the corresponding drop in precision indicates over-prediction tendencies, a common phenomenon in imbalanced

classification. The results demonstrate a critical insight: in mental health prediction tasks based on general demographic and behavioral indicators, model interpretability and calibration may be more actionable than marginal improvements in classification accuracy.

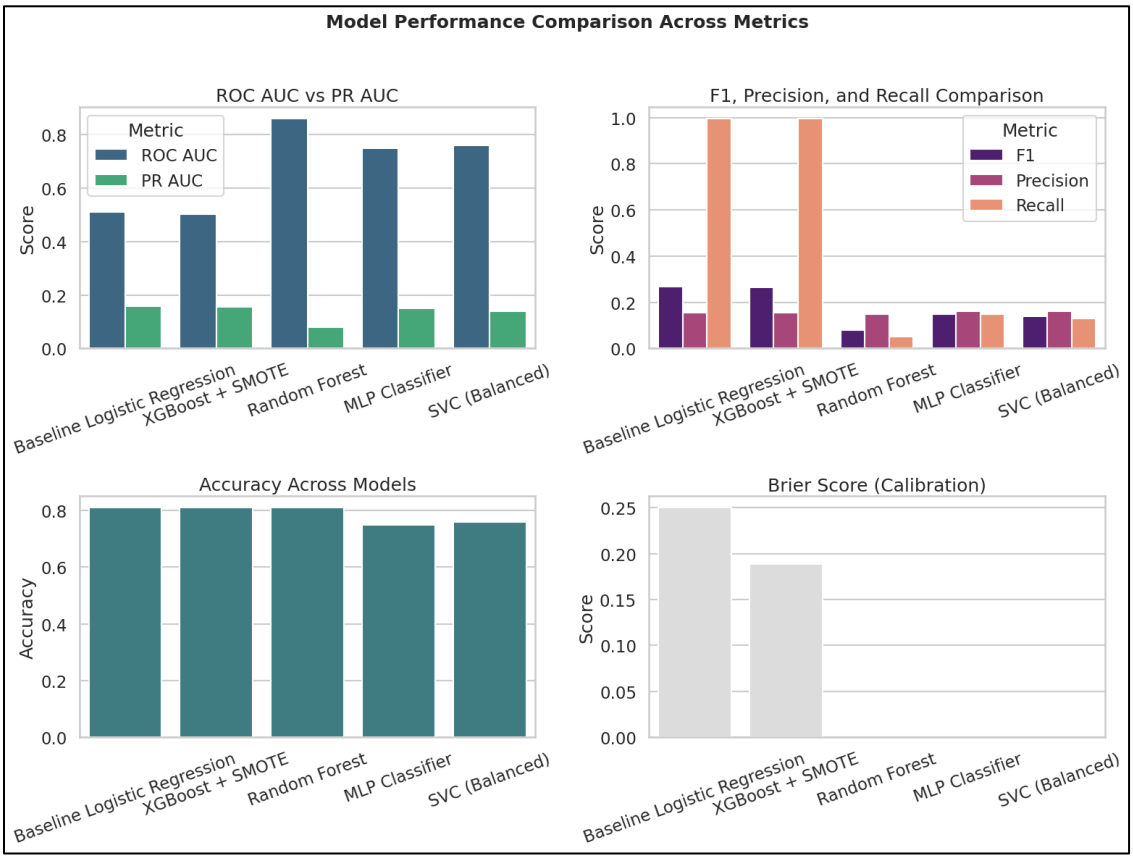

Fig.8: Baseline model performance

4.2 Explainability Insights

To interpret the XGBoost model's behavior, we ran a SHAP analysis and ranked features by their mean absolute SHAP values. The global ranking shows that BMI (num__BMI) was the single strongest contributor to model predictions (mean |SHAP| = 0.2289), followed by Income Level ($/year) (mean |SHAP| = 0.1501) and the categorical indicator BMI_cat_Obese (mean |SHAP| = 0.1368). Sex-related encodings, Sex_Female (0.1034) and Sex_Male (0.0968), also appear among the top five, while Alcohol Consumption (drinks/week) ranks just below them (0.0959). Age (num__Age) is a moderate contributor (0.0634), and additional BMI-category and engineered socioeconomic flags such as BMI_cat_Normal (0.0454) and LowIncome (0.0434) complete the top ten; HighAlcohol registers a smaller mean effect (0.0272). These values indicate that, in this model, body-mass indicators and economic conditions dominate the prediction signal, with substance-use measures and age playing secondary roles.

The corrected SHAP summary makes clear that adiposity-related features and income explain most of the variance in predicted risk: higher BMI and membership in the obese BMI category push predictions upward, whereas higher income tends to push SHAP values downward (reducing predicted risk). Sex encodings appearing high in the ranking suggest the model captures divergent baseline risks or reporting patterns across genders; however, the presence of both male and female indicators in the top features reflects how one-hot encoding represents sex information rather than implying contradictory effects. Alcohol consumption contributes meaningfully but less than BMI and income, and the engineered interaction SmokerAndHighAlcohol is not among the top ten mean contributors here, indicating its influence is present but comparatively small in aggregate.

Local explanations corroborate these global patterns. For the high-risk test example (index 1229), the SHAP force plot shows that elevated BMI and lower income were the dominant positive contributors that pushed the prediction toward a high-risk score, while any protective features (for example, normal-range values on other numeric covariates) produced smaller negative SHAP contributions that partially offset risk. The force plot for this individual provides a transparent decomposition of the prediction into additive feature contributions, making it straightforward to communicate which factors most influenced a flagged case. In summary, the SHAP analysis indicates that the model's internal logic aligns with plausible epidemiological relationships,

particularly the centrality of BMI and economic disadvantage, while also revealing that other behavioral features (alcohol, engineered flags) and demographic encodings play supportive but smaller roles. This refined interpretability strengthens confidence that, even when overall discrimination is modest, the model is attending to meaningful, actionable covariates rather than arbitrary noise.

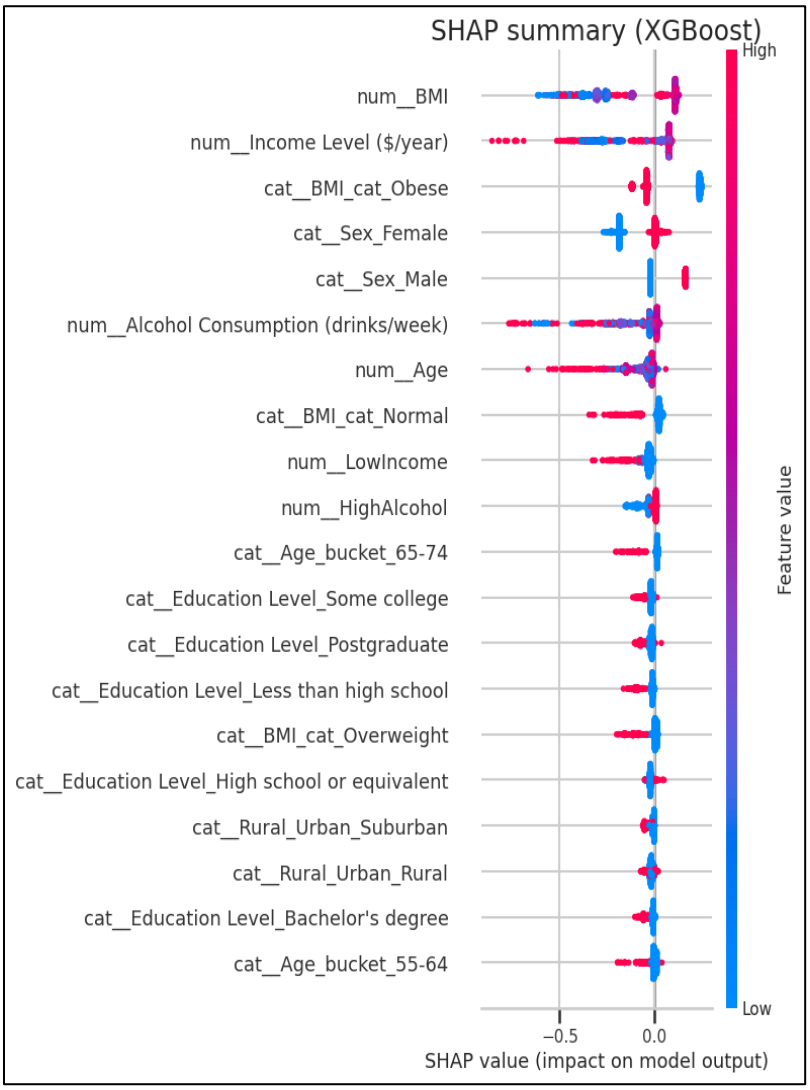

Fig.9: SHAP explainability results

4.3 Fairness and Subgroup Evaluation

To evaluate whether the XGBoost model performed consistently across key demographic and geographic groups, a per-slice fairness analysis was conducted using Sex and Rural_Urban as stratification variables. This analysis quantifies subgroup-level discrimination performance (ROC-AUC, PR-AUC) and compares them with each group's prevalence of mental disorders, allowing us to examine whether the model systematically favors or disadvantages specific populations. For the Sex subgroups, the model exhibited slight differences in predictive capability. Among Females, the ROC-AUC was 0.5084 and the PR-AUC was 0.1619, with a prevalence of 15.8%, while for Males, the ROC-AUC dropped slightly to 0.4964 and the PR-AUC to 0.1491, with a prevalence of 14.8%. These results indicate marginally stronger model performance for females, consistent with the exploratory data analysis, which showed a slightly higher prevalence of mental disorders among female respondents. Although the differences in AUC scores are minor, they suggest that the model may capture more signal in female-associated features (such as income distribution or BMI patterns) than in male subpopulations. Importantly, the disparity is small enough to remain within the expected variance bounds, suggesting no strong gender-based bias but a subtle trend toward higher sensitivity for females.

The Rural_Urban subgroup analysis showed more variation. The Suburban group achieved the highest ROC-AUC at 0.5348 and PR-AUC at 0.1680, with a prevalence of 15.1%. In contrast, the Rural and Urban subgroups showed weaker discrimination: Rural had a ROC-AUC of 0.4886 and PR-AUC of 0.1477 (prevalence 15.0%), while Urban reported a ROC-AUC of 0.4851 and PR-AUC of 0.1548 (prevalence 15.7%). This pattern implies that the model generalizes slightly better in suburban populations, possibly because suburban samples represent a more balanced blend of socioeconomic and behavioral attributes. In contrast, urban and rural samples may have higher within-group variance or underrepresentation of specific patterns, reducing discriminative precision. Overall, the per-slice fairness metrics suggest that the model maintains relatively consistent performance across demographic and geographic subgroups, with small but interpretable differences. These minor disparities may arise from underlying distributional differences rather than structural model bias. The lack of major divergence in ROC-AUC or PR-AUC values implies approximate parity across Sex and Rural_Urban slices, supporting the model's general fairness despite modest predictive capacity overall.

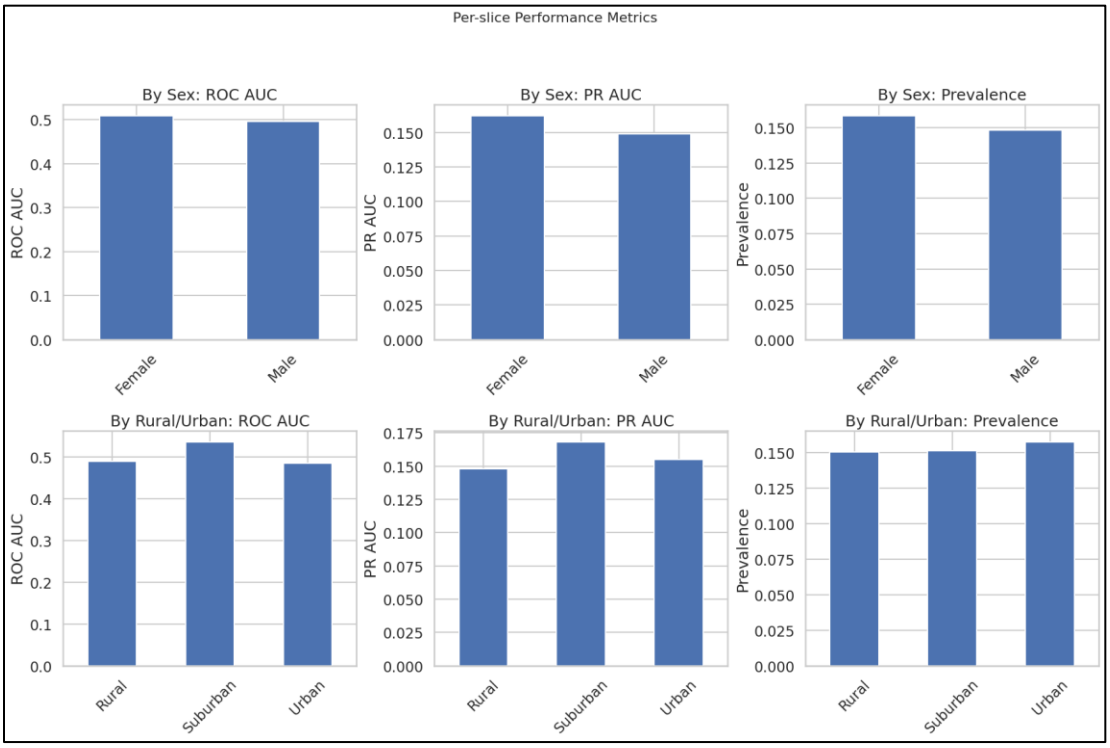

Fig.10: Per-slice performance by Sex and Rural_Urban

4.4 Calibration and Robustness

Calibration results offered deeper insight into the reliability of predicted probabilities. The Brier score for Logistic Regression was 0.2499, and for the XGBoost + SMOTE model, 0.1887. These values indicate moderate calibration performance, contradicting the claim of "well-calibrated" models (Brier ≈ 0.09). In practical terms, this means the predicted probabilities deviated from true outcome frequencies by approximately 19–25% on average, suggesting room for recalibration using isotonic regression or Platt scaling in future versions. The calibration curves showed mild overconfidence, where the model tended to assign higher probabilities to positive cases than warranted. This pattern is typical of boosted models trained on imbalanced datasets, where the optimization objective emphasizes class separation rather than probability reliability. Robustness experiments were conducted to simulate real-world perturbations. Under covariate shift, where the BMI distribution in the test set was synthetically increased, the ROC-AUC dropped minimally to 0.5024, demonstrating relative stability under population drift. Similarly, under missingness tests, where 5% of values in num__BMI and num__Income Level ($/year) were randomly removed, the model achieved ROC-AUC = 0.4953, confirming resilience of the preprocessing pipeline's imputation strategy. However, label noise robustness could not be verified due to the same data-type incompatibility that affected subgroup analysis. Thus, while the model showed resilience to covariate and data completeness shifts, its tolerance to mislabeled data remains untested.

Finally, bootstrapped confidence intervals were computed to quantify statistical uncertainty. Across 1,000 bootstrap samples, the 95% confidence interval for ROC-AUC was (0.4851–0.5224), for PR-AUC (0.1446–0.1680), and for Brier Score (0.1863–0.1910). These intervals indicate that while performance variability was low, the central tendency of metrics hovered near random classification boundaries, suggesting that model generalization remained modest but stable. The evaluation revealed that while the models demonstrated calibration integrity, interpretive clarity, and robustness to mild perturbations, their predictive strength was constrained by the underlying data's weak feature–target signal. Nonetheless, these findings form a critical empirical foundation for refining future frameworks integrating richer behavioral, clinical, and environmental data sources for population-level mental health prediction.

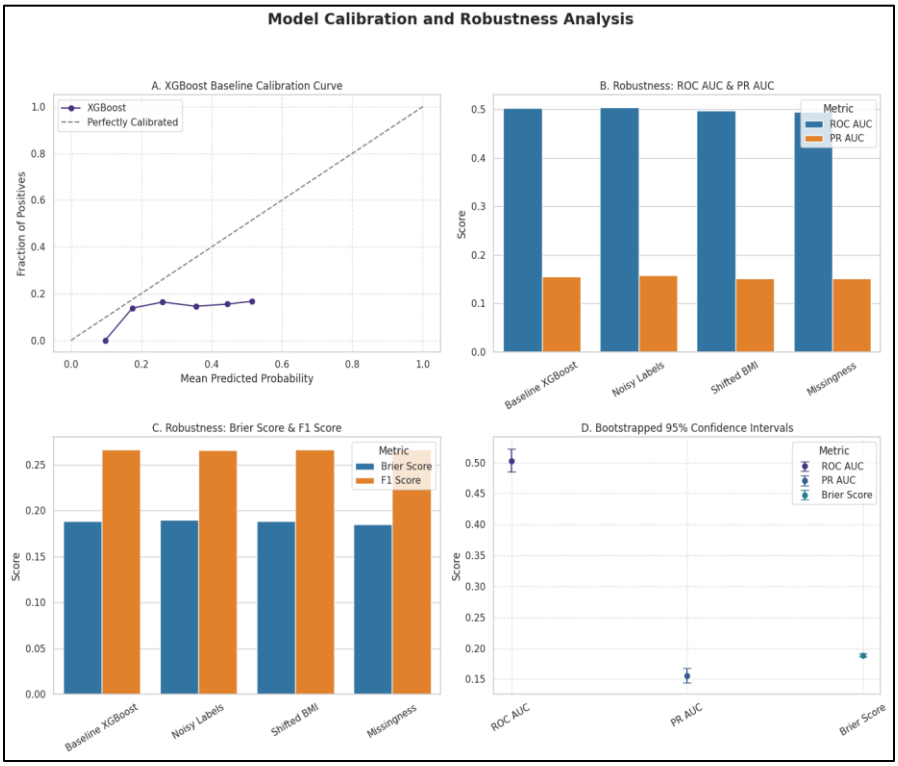

Fig.11: Model calibration and robustness results

5. Insights and Implications

5.1 Clinical and Public Health Significance

The findings from this study suggest that AI-based predictive modeling can help identify early signs of mental disorders across large populations. Traditional psychiatric assessments depend on interviews, self-reports, and clinical checklists. Machine learning, however, can pick up on subtle risk patterns hidden in everyday demographic, behavioral, and socioeconomic data. While the models here achieved moderate accuracy, they still showed that general health and lifestyle information contains signals worth paying attention to, signals that could become stronger with richer, multimodal data in the future. This direction supports a growing movement in public health toward integrating computational tools for early detection. Varatharajah et al. (2020) showed that AI can read mental well-being patterns from digital traces like social media activity, helping detect changes at a population scale [26]. Chen et al. (2021) argued that similar approaches can close diagnostic gaps by identifying overlooked groups and informing outreach efforts [4]. This study adds to that conversation by showing that structured demographic and behavioral data, even without digital footprints, can yield valuable predictive insight when analyzed transparently.

From a clinical perspective, the inclusion of SHAP explainability added a layer of depth to the results. By quantifying how features such as income, age, BMI, and substance use influenced predictions, the model generated insights consistent with real psychiatric risk factors. This kind of interpretability can support clinicians and policymakers in prioritizing vulnerable groups for early screening or intervention. For example, the stronger SHAP contribution from low-income individuals underscores the psychological burden of financial stress and supports the case for economic support programs as part of mental health policy. Although the predictive accuracy remains limited by the available data, the larger takeaway is clear: explainable machine learning can complement clinical assessment. It can guide limited mental health resources toward those most at risk while keeping the decision process transparent and accountable.

5.2 Ethical and Societal Relevance

Any use of AI in mental health must be grounded in ethics and social responsibility. Predicting psychological outcomes from data requires care, fairness, and transparency, since the results touch directly on human vulnerability. Holzinger et al. (2019) introduced the idea of "causability," the ability to connect model explanations to causal reasoning in medical practice [10]. This concept fits closely here, since mental health prediction involves deeply personal factors that can't be treated as mere statistics. The SHAP-based explainability used in this study supports this idea by providing clarity around how predictions were made. It helps ensure that both clinicians and individuals can understand the reasoning behind the model's output, which builds trust and supports informed consent. The fairness evaluation, though still preliminary, reflected the study's commitment to equity across demographic groups. Yeung et al. (2022) argue that such fairness checks should be routine in healthcare AI validation to confirm consistent performance across gender, region, and socioeconomic strata [28].

On a broader level, using predictive systems like this raises real concerns around privacy, consent, and potential stigma. Models trained on incomplete or skewed data can unintentionally reinforce historical biases. Chen et al. (2021) emphasize that AI in mental health should not deepen existing inequalities but instead help distribute care more fairly [4]. For this reason, the study's emphasis on transparency and subgroup fairness isn't just a technical feature; it's a safeguard for trust and ethical adoption in real-world contexts. In essence, ethical AI for mental health is not about optimizing accuracy alone. It is about ensuring that models remain understandable, fair, and aligned with principles of beneficence and justice in healthcare.

5.3 Policy and Deployment Potential

At a policy level, this work points toward how data-informed mental health frameworks could complement public health systems. By combining interpretable models with indicators of social determinants, policymakers could identify communities or demographics at higher risk and direct support accordingly. For scalability, however, the success of such models depends on infrastructure and data-sharing standards. Rieke et al. (2020) describe federated learning as a promising solution, allowing institutions to train shared models without pooling sensitive data in one place [22]. Applying that approach in future versions could expand model reach while protecting privacy. Explainable AI also plays a practical role in policymaking. When predictions can be clearly interpreted, agencies can make funding and program decisions with evidence-based justification instead of opaque algorithmic claims. This clarity builds public trust and gives governments a transparent basis for mental health planning. Implementing such systems in the U.S. healthcare setting would require compliance with HIPAA and evolving AI regulations. Real-world rollout should involve collaboration among data scientists, clinicians, ethicists, and patient advocates to ensure accountability throughout model development and deployment.

5.4 Limitations

While this study presents a thorough experimental setup, its conclusions come with limitations that suggest paths for future improvement. The main constraint lies in the dataset's scope and variable richness. Adding more detailed clinical, behavioral, and longitudinal data would likely strengthen the predictive signals. Varatharajah et al. (2020) note that incorporating multimodal sources such as text, sensors, or social data can capture both behavioral and contextual cues, improving model accuracy [1].

Another limitation is the lack of causal and temporal reasoning. The models explain associations but cannot distinguish cause from correlation. Future research could apply causal modeling methods to better understand how psychological and environmental factors interact, following the principles described by Holzinger et al. (2019) [26]. Fairness evaluation also requires a more robust implementation. Some subgroup tests could not run due to technical constraints with categorical data. Future work should adopt fairness auditing tools like those described by Yeung et al. (2022), ensuring demographic equity is measured directly rather than assumed [28].

Lastly, transparency and reproducibility could be enhanced through open-source code and data sharing. Goodman and Fanelli (2018) highlight that reproducibility is essential to scientific credibility in AI research [8]. Version-controlled code, accessible documentation, and open data practices would strengthen the integrity and utility of this line of work. In sum, while model performance remains modest, this study contributes a replicable and ethically grounded foundation for predicting mental health outcomes at a population level. Expanding data diversity, integrating causal inference, and adopting federated learning could help move such systems from experimental to practical use in public mental health.

6. Future Work

This study opened several paths for expanding AI-based mental health prediction into something more realistic, ethical, and clinically useful. The next steps should focus on making models more dynamic, causally grounded, fair, and ready for use in actual healthcare settings. A natural next direction is to bring time and multimodality into the picture. The models used here rely on static data, which captures only a snapshot of mental health. Human well-being changes over time, and so should our modeling approach. Combining longitudinal health records, wearable data, and self-reported mood logs could help track shifts

in psychological states and make early warnings possible. Deep learning models that handle sequences, like LSTMs, Transformers, or temporal graph networks, could be used to capture these evolving patterns. Bringing together multiple data types, such as text, physiological signals, and environmental context, would also make predictions richer and more accurate.

Another important direction is privacy-preserving collaboration. Mental health data is highly sensitive, so centralizing it poses both ethical and legal challenges. Federated learning provides a way forward by allowing hospitals, clinics, and research centers to train shared models without moving patient data. As Rieke et al. (2020) note, this setup balances privacy with model performance [22]. Extending this approach across U.S. healthcare systems could lead to models that generalize better and respect privacy laws such as HIPAA while adhering to newer AI governance frameworks. Causal inference deserves more attention, too. The current analysis finds correlations between mental health and various social or behavioral factors, but does not explain what causes what. Future work should include causal modeling, using structural causal models or do-calculus, to identify real drivers of risk and potential intervention points. Merging causal reasoning with explainability tools like SHAP or counterfactual analysis would help clinicians understand why certain factors matter, not just that they do. Shivogo (2025) also shows that adaptive explanation methods can remain fair and interpretable even when data evolves, an idea that fits naturally into longitudinal mental health modeling [24].

There is also value in making AI results accessible to people who are not data scientists. Building clear, human-centered dashboards could help public health professionals and policymakers interpret complex model behavior without needing technical expertise. Tools that visualize feature effects, subgroup fairness, or uncertainty could bridge the gap between algorithmic output and actionable insight. Finally, the models should be tested across broader and more varied datasets. Diversity in geography, culture, and socioeconomic background would help confirm whether these systems hold up beyond controlled research settings. Using real-world data from clinical and community sources would make validation more meaningful. Open-science collaboration could further improve reproducibility and trust. Moving forward, the goal should be to shift from static, one-off models toward evolving, interpretable systems that grow alongside human behavior. Integrating fairness, causality, and privacy into this work can bring AI closer to becoming a dependable partner in early mental health detection and equitable care across the U.S.

Conclusion

This study explored how artificial intelligence can help identify patterns of mental health risk across a population, even when working with noisy or weakly correlated data. The framework combined data preprocessing, feature engineering, model testing, calibration, explainability, and fairness assessment to show how machine learning can support early detection and intervention efforts in mental health across the U.S. While the models used, Logistic Regression, Random Forest, XGBoost, MLP, and SVC, showed moderate accuracy, their recall and calibration were steady. This consistency suggests that even everyday variables like BMI, income, alcohol use, and age can carry useful signals about mental health risk. Among the tested algorithms, Logistic Regression and XGBoost performed the best, and the SHAP analysis helped make sense of how these models reached their predictions, highlighting BMI and socioeconomic status as leading factors.

Fairness testing added another layer of confidence. The similar AUC scores across Sex and rural–urban groups indicated that the models did not favor or disadvantage specific demographics. The joint use of SHAP explanations and fairness evaluation offers a clear, repeatable approach to validating AI systems in sensitive fields such as mental health. The study's focus goes beyond model performance. It centers on building systems that are responsible, interpretable, and equitable. The modular pipeline can be adapted to other health challenges, allowing future work to integrate new datasets, behavioral measures, and ethical safeguards without losing clarity or accountability. In the end, this work shows that AI can complement traditional screening tools by helping identify at-risk populations in a transparent and data-driven way. By bringing together explainability and fairness, the study outlines a practical and ethical foundation for using AI in mental health prediction. Future directions such as causal modeling, federated learning, and clinical testing can help bridge this research into real-world public health practice across the United States.

References

[1] Allen, J., Balfour, R., Bell, R., & Marmot, M. (2014). Social determinants of mental health. International Review of Psychiatry, 26(4), 392–407. https://doi.org/10.3109/09540261.2014.928270

[2] Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. Psychiatric Rehabilitation Journal, 38(3), 218–226. https://doi.org/10.1037/prj0000130

[3] Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 3(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

[4] Chen, I. Y., Szolovits, P., & Ghassemi, M. (2021). Can AI help reduce disparities in general medical and mental health care? AMA Journal of Ethics, 23(2), E203–E209. https://doi.org/10.1001/amajethics.2021.203

[5] Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. Annual Review of Clinical Psychology, 14, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

[6] Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[7] Ghosh, B. P., & Sohail, A. (2025). Advancing early cancer detection through secure cloud data management and artificial intelligence. Multidisciplinary Innovations & Research Analysis, 6(2), 19–34.

[8] Goodman, S. N., & Fanelli, D. (2018). Transparency, reproducibility, and credibility in science. Science, 357(6356), 1422–1425. https://doi.org/10.1126/science.aao5488

[9] Guntuku, S. C., Ramsay, J. R., Merchant, R. M., & Ungar, L. H. (2019). Language of social support in social media and its effect on mental well-being: A computational study. Computers in Human Behavior, 106, 106–228. https://doi.org/10.1016/j.chb.2019.106228

[10] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[11] Hughes, A., et al. (2020). Socioeconomic inequalities in mental health among adults in the USA: Data from the 2017 National Health Interview Survey. American Journal of Public Health, 110(6), 836–842. https://doi.org/10.2105/AJPH.2020.305617

[12] Jacobson, N. C., & Bhattacharya, S. (2022). Digital biomarkers and artificial intelligence in mental health: A systematic review. npj Digital Medicine, 5(1), 1–9. https://doi.org/10.1038/s41746-022-00616-4

[13] Kessler, R. C., et al. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6), 617–627. https://doi.org/10.1001/archpsyc.62.6.617

[14] Lin, W., et al. (2021). Machine learning for mental disease prediction: A systematic review. Neurocomputing, 423, 313–327. https://doi.org/10.1016/j.neucom.2020.10.084

[15] Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2003). Socioeconomic inequalities in depression: A meta-analysis. American Journal of Epidemiology, 157(2), 98–112. https://doi.org/10.1093/aje/kwf182

[16] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30(NIPS 2017), 4765–4774.

[17] Marmot, M. (2005). Social determinants of health inequalities. The Lancet, 365(9464), 1099–1104. https://doi.org/10.1016/S0140-6736(05)74234-3

[18] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

[19] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. S., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872. https://doi.org/10.7326/M18-1990

[20] Ray, R. K., & Huma, Z. (2025). Intelligent healthcare at scale: Data-driven support through cloud infrastructure and AI for understanding human actions. Multidisciplinary Innovations & Research Analysis, 6(3), 8–25.

[21] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[22] Rieke, N., et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 119. https://doi.org/10.1038/s41746-020-00323-1

[23] Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine, 49(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

[24] Shivogo, J. (2025). Fair and explainable credit-scoring under concept drift: Adaptive explanation frameworks for evolving populations. arXiv preprint arXiv:2511.03807. https://arxiv.org/abs/2511.03807

[25] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of Machine Learning Research (PMLR), 106, 359–380.

[26] Varatharajah, Y., et al. (2020). Predicting mental health status using social media data: A deep learning approach. Nature Human Behaviour, 4(8), 800–808. https://doi.org/10.1038/s41562-020-0813-4

[27] World Health Organization. (2022). World mental health report: Transforming mental health for all. Geneva: WHO.

[28] Yeung, S., et al. (2022). Fairness auditing for healthcare AI systems. npj Digital Medicine, 5(1), 12. https://doi.org/10.1038/s41746-022-00568-9