

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemdiktisaintek, No.10/C/C3/DT.05.00/2025

SHAP-ALE: A novel approach to explainability in mental health prediction using RFR

Nur Alamsyah[1,*], Budiman[2], Wala Erpurini[3], Hani Fitria Rahmani[4]

[1,2]Information System, Faculty of Technology and Information, Universitas Informatika Dan Bisnis Indonesia
[3]Management, Faculty of Economics and Business, Universitas Jenderal Ahmad Yani, Bandung, Indonesia
[4]Accounting Study Program, Vocational School, IPB University, Bogor, Indonesia

ARTICLE INFO

History of the article:
Received January 2, 2026
Revised February 10, 2026
Accepted March 24, 2026

Keywords:
Explainability
SHAP
ALE
Mental Health Prediction
Machine Learning

Correspondence:

nuralamsyah@unibi.ac.id

ABSTRACT

The prediction of mental health disorders, such as depression, has become increasingly crucial as global mental health concerns continue to rise. In this study, the predictive task specifically focuses on estimating the prevalence of depressive disorders as the primary target variable, while other mental health conditions such as anxiety and schizophrenia are treated as explanatory features. While machine learning (ML) models, like Random Forest Regressor (RFR), offer high accuracy in predictions, their interpretability remains a challenge. This research introduces SHAP-ALE, an innovative hybrid explainability framework that integrates SHapley Additive exPlanations (SHAP) and Accumulated Local Effects (ALE) to address this gap. SHAP provides both global and local insights into feature contributions, while ALE visualizes feature-target relationships, mitigating bias caused by feature correlations. Using a dataset comprising various mental health disorders and demographic factors, the dataset used in this study was obtained from Kaggle and consists of 6,421 records covering multiple mental health disorder indicators and demographic attributes across different regions and years. RFR model demonstrated robust predictive performance with an R^2 score of 0.9984 and a Mean Squared Error (MSE) of 0.0016. SHAP analysis revealed that features such as schizophrenia and anxiety disorders significantly influenced predictions, while ALE identified nonlinear relationships between these features and depression prevalence. The combined insights from SHAP and ALE enhance the interpretability of the model, enabling better understanding of the complex factors underlying mental health disorders. This study highlights the contribution of SHAP-ALE as a hybrid explainability framework that integrates local and global interpretability, enabling a more comprehensive understanding of feature interactions and non-linear effects beyond the capabilities of individual methods.

1. INTRODUCTION

In recent years, mental health disorders such as depression, anxiety, and schizophrenia have emerged as critical global public health challenges, affecting millions of individuals across diverse demographics (Rotejanaprasert et al., 2024). These conditions severely impact individual well-being, disrupt social interactions, and impose significant economic burdens on healthcare systems and societies (Kirkbride et al., 2024). The World Health Organization (WHO) estimates that mental health disorders contribute to around 14% of the global disease burden, with depression being a major factor (Chen et al., 2024). Despite

the urgent need for effective intervention strategies, identifying the factors driving mental health prevalence and developing predictive tools remain complex due to the multifaceted nature of these conditions (Colizzi, 2025). These challenges emphasize the importance of early detection and targeted interventions to mitigate adverse outcomes and improve quality of life (Darko et al., 2024).

Several recent studies have highlighted the potential of SHAP and ALE as explainability tools in mental health prediction tasks. For instance, (Vu et al., 2025) applied SHAP to identify the most influential features in predicting depression from clinical and behavioral data, improving the transparency of their deep learning model. Similarly, (Chattopadhyay et al., 2025) utilized ALE plots to explore how cumulative patient history variables impacted anxiety disorder predictions, offering valuable insights for mental health professionals. Furthermore, (Chatterjee et al., 2023) demonstrated the effectiveness of combining SHAP and ALE to explain suicide risk prediction using random forest models, underlining the importance of both local and global explanations in sensitive medical applications. These studies affirm that SHAP and ALE are not only technically sound but also practically relevant in the field of mental health analytics.

The opacity of machine learning models remains a critical issue in mental health prediction, limiting their adoption in clinical decision-making (Dhanda et al., 2025). Although explainability techniques such as SHAP and LIME have been widely applied, most existing approaches focus on either local or global interpretability, without integrating both perspectives. This limitation restricts the ability to capture both individual-level feature contributions and complex non-linear relationships at a global level. Therefore, there is a need for a unified explainability framework that can provide complementary insights and improve model transparency in mental health analysis.

This study is limited to using publicly available data on the prevalence of mental health disorders and focuses on building an explainable machine learning framework rather than developing clinical interventions. The research formulates a solution by introducing a hybrid framework, SHAP-ALE, which combines SHapley Additive exPlanations (SHAP) and Accumulated Local Effects (ALE) to achieve both global and local explainability. SHAP provides granular insights into feature contributions for individual predictions, while ALE reveals global feature impacts by mitigating biases from feature correlations. Together, these approaches offer a holistic understanding of the factors influencing mental health predictions.

Explainability in machine learning, particularly in healthcare applications, has garnered increasing attention in recent years. (Parisineni & Pal, 2024) introduced LIME (Local Interpretable Model-agnostic Explanations), a widely used framework for explaining individual predictions. (Hikmawati & Alamsyah, 2024) further advanced this field with SHAP, a unified approach based on Shapley values that provides both local and global explanations. Accumulated Local Effects (ALE), proposed by (Li et al., 2025), addressed the limitations of Partial Dependence Plots (PDP) by accounting for feature correlations, offering a more robust method for global interpretability.

In the context of mental health, predictive modeling has predominantly focused on accuracy rather than explainability. Studies by (Zhai et al., 2025) and (Tran et al., 2025) utilized machine learning to predict depression and anxiety prevalence, highlighting the role of socioeconomic and demographic factors. However, these studies often lack transparent explainability frameworks, limiting their practical application

in policy and healthcare decision-making. By integrating SHAP and ALE into a single framework, this study addresses the gap in explainable AI (XAI) for mental health predictions. Unlike conventional approaches that apply SHAP and ALE independently or sequentially, this study proposes a structured integration where both methods are systematically combined to provide complementary interpretability. The SHAP–ALE framework is designed to explicitly bridge local feature attribution and global non-linear effect analysis, enabling a unified interpretation pipeline rather than separate analytical steps.

To achieve these objectives, a Random Forest Regressor (RFR) was employed as the predictive model, leveraging a dataset of mental health prevalence. The selection of Random Forest Regression is motivated by its strong capability to model complex non-linear relationships, robustness against overfitting through ensemble learning, and its compatibility with feature importance analysis, which aligns well with explainability techniques such as SHAP and ALE.. Therefore, this study aims to develop a hybrid explainability framework by integrating SHAP and ALE within a Random Forest Regression model to enhance both predictive transparency and interpretability in mental health prediction.

The findings of this study contribute to three key areas. First, it introduces an innovative hybrid framework for explainable mental health predictions, addressing the dual need for global and local interpretability. Second, it provides actionable insights into the factors influencing depression prevalence, enabling healthcare professionals and policymakers to prioritize interventions effectively. Finally, this research advances the field of explainable AI (XAI) by demonstrating the applicability of SHAP and ALE in addressing real-world public health challenges. By bridging the gap between accuracy and interpretability, this study lays the foundation for the adoption of transparent and reliable AI-driven solutions in mental health. The novelty of this study lies in the methodological integration of SHAP and ALE within a unified interpretability framework, emphasizing their complementary roles in addressing feature interaction, correlation bias, and non-linear relationships in mental health prediction models.

2. RESEARCH METHODS

The next section explains how the study was carried out and how the results can be explained. The research process involves multiple stages, beginning with data collection and pre-processing, and continuing with splitting the data into training and testing sets. A Random Forest Regressor is used as the predictive model and its performance is evaluated using standard metrics. To improve the model's interpretability, explainability techniques such as SHAP and ALE are employed. Finally, the findings are visualized. This provides actionable insights into the factors influencing mental health outcomes. Figure 1 illustrates the methodology to provide a clear overview of the research workflow.

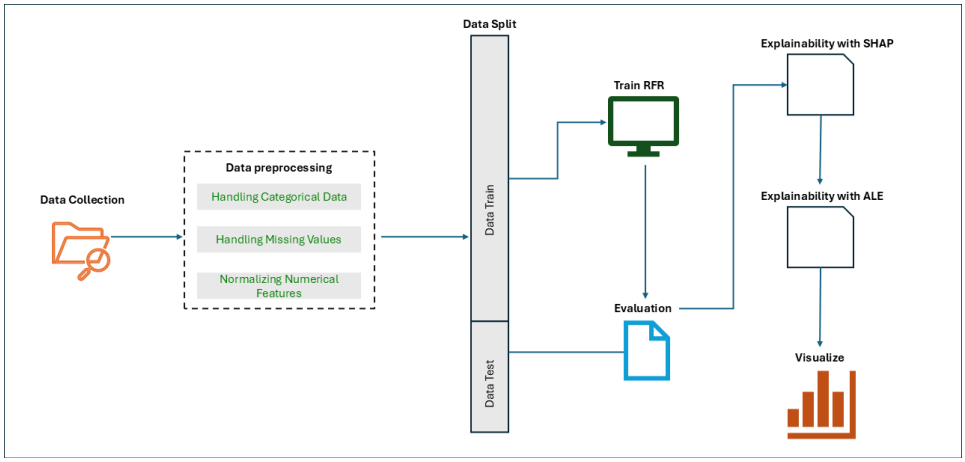


Figure 1. Our Proposed Method

2.1. Data Collection

The dataset utilized in this study was obtained from the Kaggle portal in 2024, providing a comprehensive overview of mental health disorders across various regions and years (Momeni, 2024). The dataset consists of a total of 6,421 records, covering features related to the prevalence of mental health disorders such as depression, anxiety, and schizophrenia, along with additional demographic and categorical information. These features are essential for building predictive models and interpreting the factors influencing mental health outcomes. The details of the dataset features are summarized in Table 1.

Table 1. Dataset features

Feature Name	Description	Type
Entity	Shows the country or region included in the dataset.	Categorical
Code	A unique code used to identify each region.	Categorical
Year	The year when the data was collected.	Numerical
Depressive disorders (share of population)	Percentage of the population with depressive disorders (age-standardized, both sexes).	Numerical
Anxiety disorders (share of population)	Percentage of the population with anxiety disorders (age-standardized, both sexes).	Numerical
Schizophrenia disorders (share of population)	Percentage of the population with schizophrenia (age-standardized, both sexes).	Numerical
Bipolar disorders (share of population)	Percentage of the population with bipolar disorders (age-standardized, both sexes).	Numerical
Eating disorders (share of population)	Percentage of the population with eating disorders (age-standardized, both sexes).	Numerical

2.2. Data Preprocessing

To ensure the dataset is suitable for machine learning modeling, a series of preprocessing steps were applied. These steps are essential to handle inconsistencies in the data and to standardize the format for

effective model training (Alamsyah, Yoga, et al., 2024). The preprocessing process includes three main stages: handling categorical data, addressing missing values, and normalizing numerical features.

2.2.1. Handling Categorical Data

The dataset contains categorical features, such as Entity and Code, which cannot be directly processed by machine learning models. These features were encoded using Label Encoding, where unique categorical values were transformed into numerical representations. For example, each country or region in the Entity column was assigned a unique integer value.

2.2.2. Handling Missing Values

Missing values were identified in the Code column, which represents unique region identifiers. To address this issue, missing values were replaced with the placeholder "Unknown." However, this approach may introduce bias, especially when categorical variables are correlated. More advanced techniques, such as imputation based on related features or removal of incomplete records, could provide more reliable preprocessing and are suggested for future research.

2.2.3. Normalizing Numerical Features

The numerical features in the dataset, such as Year, Depressive disorders (share of population), and other prevalence-related metrics, were normalized using StandardScaler. While normalization improves numerical stability during model training, it may reduce the interpretability of the target variable, particularly when transformed values include negative ranges. Therefore, interpretation of the results should consider the scaled representation rather than the original units. This method standardizes the features to have a mean of 0 and a standard deviation of 1, ensuring that all features are on the same scale. Normalization is crucial to prevent features with larger scales from dominating the model training process.

2.3. Data Split

The dataset was split into two subsets after the preprocessing steps were completed: a training set and a testing set. The machine learning model was trained on the training set, which comprised 0.8 of the data. The remaining 0.2 of the data formed the testing set, which was used to evaluate the model's performance on unseen data (Biggs et al., 2024). This splitting strategy ensures that the model's ability to generalize is assessed effectively, preventing overfitting and providing a realistic evaluation of its predictive accuracy. The data split was performed randomly to maintain the representativeness of both subsets while ensuring that the distribution of features and target values remained consistent across the splits. In addition to the train–test split, cross-validation is recommended as a robust validation strategy to further assess model stability and generalization. This will be considered in future work to enhance evaluation reliability.

2.4. Random Forest Regression Train

The study used the random forest regression algorithm as its predictive model. RFR is a machine learning method that builds multiple decision trees during training. These trees are then combined to produce more accurate and robust results (Alamsyah, Kurniati, et al., 2024). This approach reduces the risk of overfitting and enhances generalization by averaging the outcomes of individual decision trees.

RFR works by minimizing the Mean Squared Error (MSE) for the target variable (y), which represents the prevalence of depressive disorders in this study. The prediction ($\hat{y}$) for a given input (X) is computed as the average of predictions from all decision trees in the ensemble:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \quad (1)$$

$\hat{y}$ represents the final output generated by the Random Forest model, Tdenotes the total number of trees in the ensemble, and $f_t(X)$ corresponds to the prediction produced by the t-th tree (f_t) for a given input X. Each tree is built using a randomly sampled portion of the training data (bootstrap sampling) and considers a random subset of features when performing splits. This process introduces variation among the trees, which helps enhance the overall predictive performance.

The model was trained using the training portion of the preprocessed dataset. Several hyperparameters, including the number of trees (T), maximum tree depth, and the number of features evaluated at each split, were adjusted to achieve optimal performance. The resulting Random Forest model is used to estimate the prevalence of depressive disorders based on the input variables and serves as the foundation for further interpretability analysis using SHAP and ALE.

2.5. Evaluation

The performance of the Random Forest Regression (RFR) model was evaluated using two key metrics:

2.5.1. Mean Squared Error (MSE) :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

(n) is the number of samples in the test dataset, (y_i) represents the actual value of the target variable for the (i)-*th* sample and ($\hat{y}_i$)denotes the predicted value for the (i)-*th* sample. This metric calculates the average squared difference between the actual values (y_i) and the predicted values ($\hat{y}_i$). A lower MSE value indicates better predictive accuracy, as it reflects how close the predictions are to the actual values.

2.5.2. R-Squared (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

n denotes the total number of samples in the test dataset, y_i is the actual value of the target variable for the i-th observation, and $\hat{y}_i$ is the predicted value for that observation. Meanwhile, $\bar{y}$represents the average value of all actual target values.

2.6. SHapley Additive exPlanations

SHAP (Shapley Additive exPlanations) is used to interpret the Random Forest Regression model by measuring the contribution of each feature to individual predictions (Timilsina et al., 2024). Built on the concept of cooperative game theory, SHAP estimates the contribution of a feature $i$by calculating its

average effect across all possible combinations of feature subsets. The SHAP value (ϕ_i) for a feature $i$is formally defined as follows:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \cdot (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (4)$$

S denotes a subset of features that does not include feature i, and Frepresents the complete set of features. The term $f(S)$refers to the model's prediction when only the features in $S$are used, while $f(S \cup \{i\})$indicates the prediction after adding feature $i$to the subset. The combinatorial factors, such as $|S|!$, $(|F| - |S| - 1)!$, and $|F|!$, are applied to ensure a fair distribution of contributions by considering all possible feature combinations.

The prediction for any instance, $(\hat{y})$, can be decomposed as the sum of the base value (ϕ_0) and the SHAP values for all features:

$$\hat{y} = \phi_0 + \sum_{i=1}^{|F|} \phi_i \quad (5)$$

Here, $\phi_0$represents the base value, which corresponds to the average prediction across all samples, while each $\phi_i$indicates the contribution of feature $i$to the final prediction. SHAP offers both global and local interpretability. At the global level, it summarizes feature contributions across all observations, typically through summary plots that rank features based on their overall influence. At the local level, SHAP explains individual predictions using visualization tools such as force plots, which illustrate how each feature contributes relative to the base value.

By combining global insights with instance-level explanations, SHAP helps balance predictive performance and interpretability. This capability makes it a valuable approach for identifying key factors in mental health prediction, improving transparency, and supporting more informed decision-making in healthcare contexts.

2.7. Accumulated Local Effects

Accumulated Local Effects (ALE) was utilized as a complementary method to explain the impact of individual features on the target variable in the Random Forest Regression model (Fisher et al., 2024). ALE focuses on providing global interpretability by quantifying the average change in the model's predictions when a feature is varied, while other features are held constant. This method mitigates the issue of feature correlation, which often biases traditional techniques like Partial Dependence Plots (PDP). The ALE value for a feature (x_j) is computed as:

$$\mathrm{ALE}(x_j) = \int_{z_{j,\min}}^{z_j} E\left[f(x_j = z_j) - f(x_j = z_j') \mid x_{\setminus j} \right] dz_j \quad (6)$$

(z_j) is a specific value of the feature(x_j), and $(z_{j,\min})$ is the minimum value of (x_j), $(f(x_j))$ represents the model's prediction when feature (x_j) takes a specific value, $(x_{\setminus j})$ denotes all other features except (x_j) and The expectation (E) is taken over the marginal distribution of the other features.

This integral-based approach calculates the average change in the model's prediction as the feature (x_j) transitions from its minimum value to any specified value(z_j). Unlike SHAP, which provides local insights, ALE focuses on capturing the global effect of a feature by averaging over the data distribution. ALE plots visualize the relationship between a feature and the target variable while accounting for potential feature interactions and correlations. The plots show how changes in a feature's value influence the target variable, highlighting whether a feature has a positive, negative, or neutral impact. By combining ALE with SHAP, this study captures both global and local interpretability, ensuring a comprehensive understanding of the model's behavior. ALE's robustness against feature correlation makes it a valuable tool for examining complex relationships in mental health prediction models.

2.8. Visualize

The final step in the methodology involves visualizing the results from the explainability techniques, SHAP and ALE, to enhance interpretability and provide actionable insights (Alamsyah et al., 2023). Visualization for SHAP includes the summary plot, which ranks features based on their average SHAP values, offering a global perspective on feature importance across all predictions. Additionally, force plots are used for local interpretability, breaking down individual predictions into feature contributions, highlighting how specific features increase or decrease the predicted value relative to the base value. For ALE, visualizations include ALE plots, which depict the cumulative effect of each feature on the model's predictions while accounting for feature interactions and correlations. These plots reveal whether a feature's relationship with the target variable is positive, negative, or non-linear. Together, these visualizations provide a comprehensive understanding of the model's behavior, bridging global and local interpretability. They are crucial for communicating findings to both technical and non-technical stakeholders, ensuring transparency and supporting informed decision-making in mental health analysis.

3. RESULTS AND DISCUSSION

3.1. Results

The performance of the Random Forest Regression model was evaluated using Mean Squared Error (MSE) and R-squared (R^2). The model achieved a low MSE of 0.0016, indicating that the average squared difference between predicted and actual values is very small. The high R^2 value (0.9984) suggests strong predictive capability; however, it may also indicate the possibility of overfitting. This outcome is likely influenced by the structured and aggregated characteristics of the dataset.

To mitigate this issue, careful preprocessing and proper data splitting strategies were implemented to prevent data leakage and ensure a fair evaluation of the model's generalization ability. While no data leakage was observed, the high R^2 score may still be affected by the dataset structure. Therefore, additional validation using cross-validation techniques and external datasets is recommended to further assess the robustness of the model.

Moreover, the R^2 value of 0.9984 indicates that the model can explain approximately 99.84% of the variance in the target variable, highlighting its high predictive accuracy and generalization potential. To provide further insight into the model's performance, Figure 2 illustrates a comparison between actual and

predicted values for the first 100 samples in the test dataset. The blue bars represent the true values (i.e., the actual prevalence of depressive disorders).

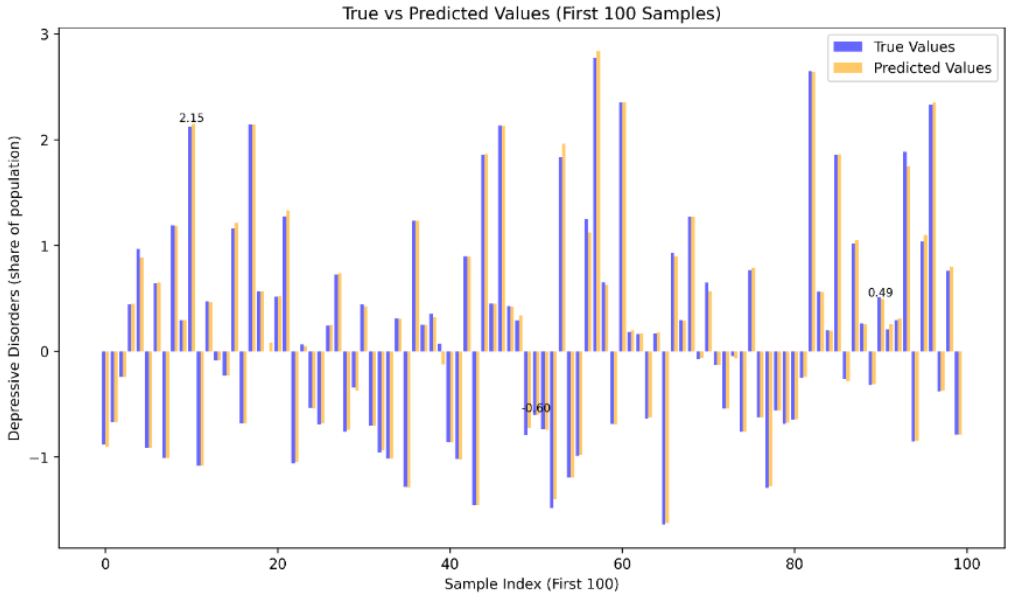


Figure 2. Comparison Between the Actual and Predicted Values

The figure shows a strong alignment between the actual and predicted values, with only minor discrepancies visible in a few samples. These deviations are relatively small, emphasizing the model's reliability. Notably, the visualization captures the entire range of predictions, including both positive and negative values of depressive disorder prevalence, indicating the model's capability to handle a diverse dataset effectively. It should be noted that negative values arise from the normalization process and do not represent actual negative prevalence, but rather standardized deviations from the mean. The inclusion of specific annotations for some key points, such as maximum and minimum deviations, further highlights the model's strengths and provides additional context for interpretation. Overall, this visualization reinforces the quantitative evaluation metrics, illustrating the model's excellent performance in predicting the prevalence of depressive disorders. The close match between the actual and predicted values underscores the utility of the RFR model for mental health predictions, providing a solid foundation for further explainability analysis using SHAP and ALE.

Figure 3 presents the SHAP summary plot, which illustrates the global importance of each feature in predicting depressive disorders. The visualization has been improved by adjusting axis labels, font size, and layout to enhance readability and ensure that feature names are clearly visible. The horizontal axis represents the mean absolute SHAP values, indicating the average impact of each feature on the model's predictions. Features are ranked in descending order of importance, with the most influential features appearing at the top. The results reveal that "Schizophrenia disorders (share of population)" has the highest SHAP value, signifying its dominant role in influencing the predictions. This is followed by "Anxiety disorders", "Bipolar disorders", and "Eating disorders", which also show substantial contributions. These features are directly related to mental health conditions, reinforcing their significance in the model's ability to predict depressive disorders.

Conversely, features such as "Entity", "Year", and country codes like "Code_GRL", "Code_UGA", and others exhibit lower SHAP values, indicating minimal influence on the predictions. This ranking provides valuable insights into which variables are critical for accurate predictions and which ones contribute less. The summary plot highlights the average contribution of each feature across all predictions, offering a global perspective on the model's decision-making process. By focusing on the features with the highest SHAP values, stakeholders can better understand the driving factors behind depressive disorders, enabling targeted interventions and resource allocation in mental health initiatives.

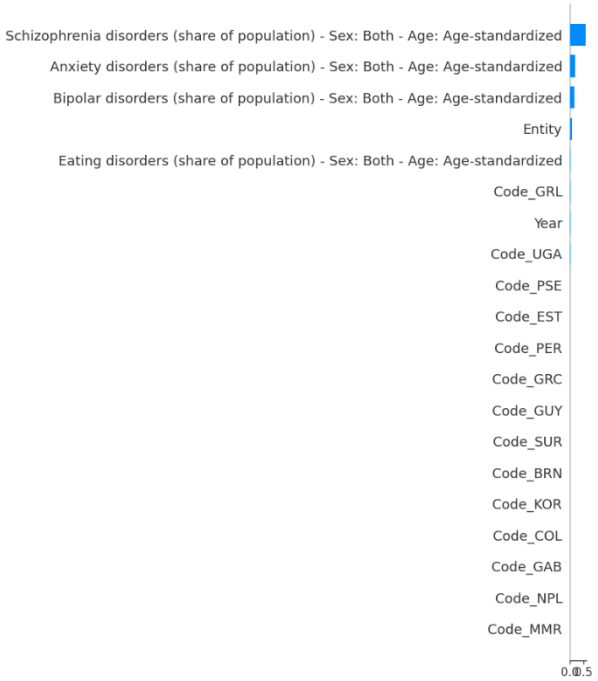


Figure 3. SHAP Summary Plot showing feature importance based on mean absolute SHAP values. Higher values indicate greater contribution of features to the model prediction.

Figure 4 illustrates the Accumulated Local Effects (ALE) plot for the feature "Entity", providing insights into its global impact on the predicted prevalence of depressive disorders. It is important to note that "Entity" is a categorical variable encoded numerically, and the encoded values do not represent ordinal relationships. Therefore, the interpretation of ALE results for this feature should be approached with caution, as the apparent patterns may not reflect meaningful numerical relationships. The ALE plot represents the average change in the model's predictions as the feature "Entity" varies, while other features are held constant. The vertical axis corresponds to the ALE values, indicating the direction and magnitude of the feature's effect on the predictions. The horizontal axis represents the transformed values of the feature "Entity."

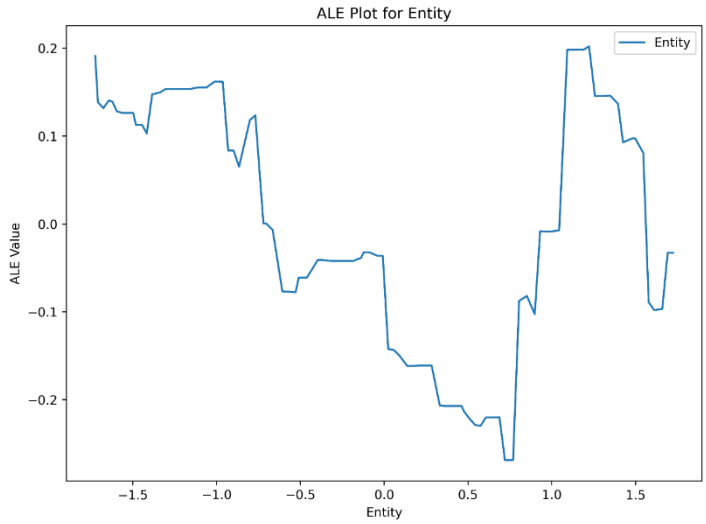


Figure 4. ALE Plot For The Feature "Entity"

The plot shows that as the "Entity" values transition, the ALE values fluctuate, highlighting non-linear relationships. Initially, the ALE values are positive, suggesting that certain values of "Entity" contribute to an increase in the predicted prevalence of depressive disorders. However, as the feature values progress, the ALE values become negative, indicating that specific ranges of "Entity" contribute to a decrease in the model's predictions. Sharp changes, such as the steep increase observed near 1.0, reveal critical regions where "Entity" has a significant impact on the predictions. This visualization emphasizes how variations in "Entity" influence the model's outputs globally, providing actionable insights into its role in the prediction process. By interpreting these trends, stakeholders can better understand the importance and interaction of "Entity" within the context of the predictive model for depressive disorders.

Figure 5 presents the Accumulated Local Effects (ALE) plot for the feature "Year", illustrating its global impact on the predicted prevalence of depressive disorders. The horizontal axis represents the normalized values of the "Year" feature, while the vertical axis shows the ALE values, which quantify the average change in the model's predictions as the "Year" feature varies. The plot reveals a clear downward trend in the ALE values as the "Year" feature increases. This indicates that higher values of "Year" contribute to a reduction in the predicted prevalence of depressive disorders. Initially, the ALE values are positive, suggesting that earlier years have a slightly positive effect on the predictions. However, as the "Year" progresses, the ALE values transition to negative, highlighting that more recent years are associated with lower predicted prevalence rates.

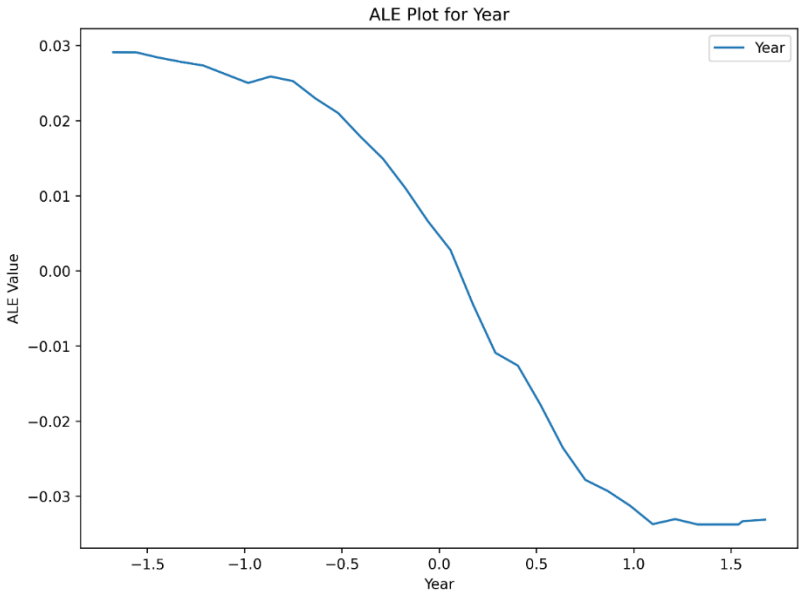


Figure 5. ALE plot for the feature "Year"

The smooth nature of the curve suggests a consistent and linear relationship between the "Year" feature and the target variable. This insight aligns with potential real-world trends, where improvements in mental health awareness and interventions over time may contribute to reduced depressive disorder prevalence. By analyzing this ALE plot, stakeholders can gain a better understanding of how temporal factors, represented by "Year," influence the model's predictions and potentially inform policy or intervention strategies in mental health contexts.

Figure 6 illustrates the Accumulated Local Effects (ALE) plot for the feature "Schizophrenia disorders (share of population)", highlighting its global impact on the predicted prevalence of depressive disorders. The horizontal axis represents the normalized values of the schizophrenia disorder prevalence, while the vertical axis displays the ALE values, quantifying the average change in predictions as the schizophrenia feature varies. The plot reveals a complex, non-linear relationship between the schizophrenia disorder prevalence and the predicted depressive disorder prevalence. At the beginning of the range, an increase in schizophrenia prevalence leads to a significant positive impact on depressive disorder predictions, as shown by the steep rise in ALE values. However, as the schizophrenia prevalence increases further, the ALE values decline sharply, indicating a transition to a negative impact on predictions.

Beyond this sharp decline, the ALE values stabilize with minor fluctuations, and at higher schizophrenia prevalence values, the impact becomes slightly positive again. This pattern suggests that the relationship between schizophrenia prevalence and depressive disorder predictions is influenced by interactions with other features or varying dynamics within the dataset.

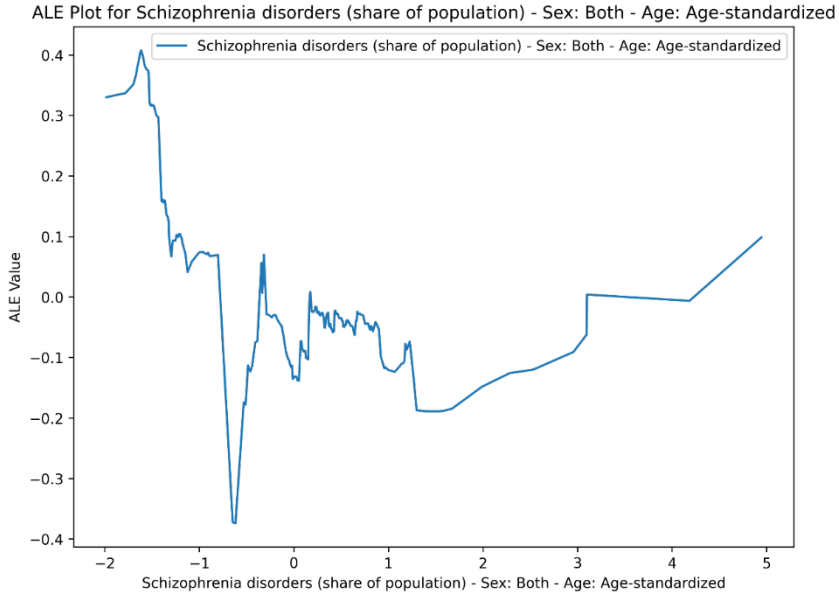


Figure 6. Plot For the Feature "Schizophrenia Disorders (share of population)"

This ALE plot provides valuable insights into the model's global behaviour, demonstrating how changes in the schizophrenia prevalence affect depressive disorder predictions. The non-linear nature of the relationship underscores the importance of accounting for feature interactions when interpreting predictive models, particularly in complex domains like mental health analysis.

3.2. Discussion

The integration of SHAP and ALE within the proposed framework addresses the inherent trade-off between model accuracy and interpretability in predictive modelling. While complex models such as Random Forest often provide high predictive accuracy, they are typically viewed as black-box models, making it difficult to understand the rationale behind their predictions. SHAP offers local interpretability by quantifying the contribution of each feature to individual predictions, allowing stakeholders to trace specific outcomes (Abioye et al., 2025). On the other hand, ALE complements this by providing a global view of how changes in feature values cumulatively affect the model output. Together, SHAP and ALE bridge the gap by retaining the accuracy of advanced ensemble models while offering both granular and generalized insights. This dual perspective enhances user trust, supports decision-making in sensitive fields such as mental health, and allows the model to be both powerful and transparent.

This study introduced a novel approach to explaining mental health predictions by integrating SHAP and ALE with a Random Forest Regression (RFR) model. The methodology was designed to address the need for interpretability in machine learning models for mental health analysis. In comparison to existing studies, this research offers distinct advantages and contributions, as highlighted below. The study by (Enkhbayar et al., 2025) employed Random Forest models for depression prediction and achieved an accuracy of 99.30%. Similar to our approach, their research utilized SHAP for model explainability, emphasizing global and local feature contributions. However, our study extends this framework by incorporating ALE, which addresses the limitations of SHAP in accounting for feature correlations. By combining SHAP and ALE, our approach provides a more comprehensive understanding of feature interactions and their cumulative impact on predictions, which was not explored in their work.

Another study by (Shen et al., 2025) focused on explainability frameworks using SHAP and LIME for healthcare applications. While their study demonstrated the effectiveness of SHAP in enhancing transparency, it lacked an analysis of non-linear relationships and feature interactions, which are critical in complex domains like mental health. Our study fills this gap by leveraging ALE to visualize the global effects of features, such as "Schizophrenia disorders" and "Year," highlighting non-linear patterns that significantly influence depressive disorder predictions.

In the research by (Ballester et al., 2021), the authors applied explainability techniques to suicide risk assessments using SHAP. Their findings highlighted the importance of feature importance rankings in clinical applications. Similarly, our SHAP summary plots identified key contributors to depressive disorder predictions, such as "Schizophrenia disorders" and "Anxiety disorders." However, our inclusion of ALE provided additional insights into the nuanced relationships between these features and the target variable, offering a more detailed understanding of the model's behaviour.

Compared to studies that relied solely on SHAP or other interpretability methods, our combined approach demonstrates enhanced robustness and interpretability. The results indicate that SHAP is well-suited for identifying critical features and providing localized explanations, while ALE complements these insights by capturing global trends and addressing feature correlation challenges. Future work will include benchmarking SHAP–ALE against other explainability methods such as LIME and PDP to further validate its effectiveness.

Table 1. Dataset Features

Aspect	This Study (SHAP-ALE)	Study by (Enkhbayar et al., 2025)	Study by (Shen et al., 2025)	Study by (Ballester et al., 2021)
Explainability Methods	SHAP + ALE (Hybrid Framework)	SHAP	SHAP + LIME	SHAP
Feature Interaction Handling	Yes (via ALE)	No	No	No
Global Interpretability	SHAP Summary Plot + ALE Plot	SHAP Summary Plot	SHAP Summary Plot	SHAP Summary Plot
Local Interpretability	SHAP Force Plot	SHAP Force Plot	LIME	SHAP Force Plot
Analysis of Non-Linear Trends	Yes (via ALE)	No	No	No
Domain Application	Depressive Disorder Prediction	Depression Prediction	Healthcare Applications	Suicide Risk Assessment
Key Contributions	Integration of SHAP and ALE for comprehensive interpretability	High accuracy and SHAP-based insights	Transparency in healthcare predictions	Feature ranking for clinical use

Table 1 highlights the key contributions of this study compared to previous works. The hybrid integration of SHAP and ALE uniquely allows this research to provide both local and global explainability, along with support for non-linear and interaction effects. This combination sets the foundation for a more interpretable and reliable machine learning application in the sensitive and complex domain of mental health prediction.

4. CONCLUSIONS AND RECOMMENDATIONS

This study proposed a hybrid framework combining SHAP and ALE for explainable mental health prediction using a Random Forest Regression (RFR) model. The results demonstrate the effectiveness of this approach in providing both local and global interpretability, addressing the limitations of relying solely on either method. Key features, such as "Schizophrenia disorders" and "Anxiety disorders," were identified as significant contributors to the prediction of depressive disorders. By leveraging SHAP's localized explanations and ALE's global feature effects, the study provides a comprehensive understanding of feature interactions and non-linear relationships, offering valuable insights into the underlying factors influencing mental health outcomes. The findings highlight the benefits of integrating SHAP and ALE for explainability, showcasing its potential for broader applications in various domains, including healthcare and social sciences. The proposed framework enhances transparency, making predictions more interpretable and trustworthy for stakeholders. Furthermore, the methodology is not limited to depressive disorders and can be adapted for other applications requiring explainability and reliability.

To further advance this research, several recommendations are proposed. First, future studies could explore the application of the SHAP-ALE framework to longitudinal datasets, enabling the prediction of mental health trends over time and identifying temporal patterns. Second, incorporating real-time data sources, such as social media or wearable devices, could enhance the model's ability to capture dynamic changes in mental health conditions. Third, developing interactive and user-friendly visualizations for SHAP and ALE outputs could make the explainability framework more accessible to clinicians, policymakers, and other stakeholders in mental health. Additionally, applying the framework to diverse machine learning models, such as gradient boosting or deep learning, could validate its generalizability and effectiveness across different predictive approaches. Finally, future research could delve into the theoretical foundations of combining SHAP and ALE, potentially leading to the development of new explainability methods that optimize interpretability and computational efficiency.

In conclusion, the hybrid SHAP-ALE framework introduced in this study paves the way for more transparent, reliable, and impactful applications of machine learning in mental health research and beyond. By addressing current limitations and providing actionable insights, this methodology has the potential to advance both theoretical and practical applications of explainable AI in critical domains.

REFERENCES

Abioye, S. O., Babatunde, Y. O., Abikoye, O. A., Shaibu, A. N., & Bankole, B. J. (2025). Optimized machine learning algorithms with SHAP analysis for predicting compressive strength in high-performance concrete. *AI in Civil Engineering*, *4*(1), 16.

Alamsyah, N., Kurniati, A. P., & others. (2023). A Novel Airfare Dataset To Predict Travel Agent Profits Based On Dynamic Pricing. *2023 11th International Conference on Information and Communication Technology (ICoICT)*, 575–581.

Alamsyah, N., Kurniati, A. P., & others. (2024). Event Detection Optimization Through Stacking Ensemble and BERT Fine-Tuning for Dynamic Pricing of Airline Tickets. *IEEE Access*. Vol. 12

Alamsyah, N., Yoga, T. P., Budiman, B., & others. (2024). IMPROVING TRAFFIC DENSITY PREDICTION USING LSTM WITH PARAMETRIC ReLU (PReLU) ACTIVATION. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *9*(2), 154–160.

Ballester, P. L., Cardoso, T. de A., Moreira, F. P., da Silva, R. A., Mondin, T. C., Araujo, R. M., Kapczinski, F., Frey, B. N., Jansen, K., & de Mattos Souza, L. D. (2021). 5-year incidence of suicide-risk in youth: A gradient tree boosting and SHAP study. *Journal of Affective Disorders*, *295*, 1049–1056.

Biggs, F., Schrab, A., & Gretton, A. (2024). MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems*, *36*.

Chatterjee, S., Mishra, J., Sundram, F., & Roop, P. (2023). Towards personalised mood prediction and explanation for depression from biophysical data. *Sensors*, *24*(1), 164.

Chattopadhyay, S., Barman, S., & Lakshmi, D. (2025). The Role of Explainable AI for Healthcare 5.0: Best Practices, Challenges, and Opportunities. *Edge AI for Industry 5.0 and Healthcare 5.0 Applications*, 45–80.

Chen, Q., Huang, S., Xu, H., Peng, J., Wang, P., Li, S., Zhao, J., Shi, X., Zhang, W., Shi, L., & others. (2024). The burden of mental disorders in Asian countries, 1990–2019: An analysis for the global burden of disease study 2019. *Translational Psychiatry*, *14*(1), 167.

Colizzi, M. (2025). Ten priorities that are shaping the future of mental health prevention. *Discover Mental Health*, *5*(1), 140.

Darko, A. P., Antwi, C. O., Adjei, K., Zhang, B., & Ren, J. (2024). Predicting determinants influencing user satisfaction with mental health app: An explainable machine learning approach based on unstructured data. *Expert Systems with Applications*, *249*, 123647.

Dhanda, S. S., Panwar, D., Lin, C.-C., Sharma, T. K., Rastogi, D., Bindewari, S., Singh, A., Li, Y.-H., Agarwal, N., & Agarwal, S. (2025). Advancement in public health through machine learning: A narrative review of opportunities and ethical considerations. *Journal of Big Data*, *12*(1), 154.

Enkhbayar, D., Ko, J., Oh, S., Ferdushi, R., Kim, J., Key, J., & Urtnasan, E. (2025). Explainable Artificial Intelligence Models for Predicting Depression Based on Polysomnographic Phenotypes. *Bioengineering*, *12*(2), 186.

Fisher, J., Allen, S., Yetman, G., & Pistolesi, L. (2024). Assessing the influence of landscape conservation and protected areas on social wellbeing using random forest machine learning. *Scientific Reports*, *14*(1), 11357.

Hikmawati, E., & Alamsyah, N. (2024). Supervised Learning for Emotional Prediction and Feature Importance Analysis Using SHAP on Social Media User Data. *Ingénierie Des Systèmes d'Information*, *29*(6).

Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., & others. (2024). The social determinants of mental health and disorder: Evidence, prevention and recommendations. *World Psychiatry*, *23*(1), 58.

Li, C., Huang, Z., Hao, H., Shen, Z., Zhao, G., Xu, B., & Liu, H. (2025). Interpretable machine learning model of effective mass in perovskite oxides with cross-scale features. *Journal of Materiomics*, *11*(1), 100848.

Momeni, M. (2024). *Mental Health* [Dataset]. https://www.kaggle.com/datasets/imtkaggleteam/mental-health/data

Parisineni, S. R. A., & Pal, M. (2024). Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. *International Journal of Data Science and Analytics*, *18*(4), 457–466.

Rotejanaprasert, C., Thanutchapat, P., Phoncharoenwirot, C., Mekchaiporn, O., Chienwichai, P., & Maude, R. J. (2024). Investigating the spatiotemporal patterns and clustering of attendances for mental health services to inform policy and resource allocation in Thailand. *International Journal of Mental Health Systems*, *18*(1), 19.

Shen, A., Sun, J., Chen, X., & Gao, X. (2025). A data-centric and interpretable EEG framework for depression severity grading using SHAP-based insights. *Journal of NeuroEngineering and Rehabilitation*, *22*(1), 116.

Timilsina, M. S., Sen, S., Uprety, B., Patel, V. B., Sharma, P., & Sheth, P. N. (2024). Prediction of HHV of fuel by Machine learning Algorithm: Interpretability analysis using Shapley Additive Explanations (SHAP). *Fuel*, *357*, 129573.

Tran, T., Tan, Y. Z., Lin, S., Zhao, F., Ng, Y. S., Ma, D., Ko, J., & Balan, R. (2025). Exploring key factors influencing depressive symptoms among middle-aged and elderly adult population: A machine learning-based method. *Archives of Gerontology and Geriatrics*, *129*, 105647.

Vu, T., Yamamoto, M., Tay, J. T., Watanabe, N., Kuriya, Y., Oya, A., Tran, P. N. H., Araki, M., & others. (2025). Prediction of depressive disorder using machine learning approaches: Findings from the NHANES. *BMC Medical Informatics and Decision Making*, *25*(1), 1–12.

Zhai, Y., Zhang, Y., Chu, Z., Geng, B., Almaawali, M., Fulmer, R., Lin, Y.-W. D., Xu, Z., Daniels, A. D., Liu, Y., & others. (2025). Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students. *Journal of Counseling & Development*, *103*(1), 110–125.