Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER
PUBLISHER

Research Article

A Hybrid Mathematical Framework Combining Logistic Regression and Neural Networks with Explainable AI Techniques for Mental Health Prediction

Arsalan Humayun[1,2], **Mohamad Arif Bin Awang Nawi**[3*], **Muhammad Ilyas Siddiqui**[4], **Russell Kabir**[5], **Abdulhafeez Babalola**[6]

[1] Bibi Aseefa Dental College, Shaheed Mohtarma Benazir Bhutto Medical University, Larkana, Pakisan
[2] School of Dental Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, Kelantan, 16150, Malaysia
[3] Biostastistics Unit, School of Dental Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, Kelantan, 16150, Malaysia
[4] Department of Liaquat University of Medical & Health Sciences, Jamshoro, Pakistan
[5] Department of Anglia Ruskin University, East Rd, Cambridge CB1 1PT, United Kingdom
[6] Biomedicine Programme, School of Health Sciences, Health Campus, Universiti Sains Malaysia, Kubang Kerian, Kelantan, 16150, Malaysia
E-mail: mohamadarif@usm.my

Received: 31 July 2025; **Revised:** 24 August 2025; **Accepted:** 9 September 2025

Abstract: Accurate prediction of mental health outcomes is vital for early intervention and effective resource allocation, yet existing methods often struggle to balance predictive accuracy with interpretability, a key requirement in clinical and policy settings. This study addresses this dual challenge by introducing a hybrid mathematical framework that combines Multiple Logistic Regression (MLR), known for its transparency, is used with a Multilayer Perceptron (MLP), which is recognized for its ability to capture complex non-linear patterns. Explainability is further enhanced through the incorporation of SHapley Additive exPlanations (SHAP) for global feature attribution and Local Interpretable Model-Agnostic Explanations (LIME) for case-specific interpretive clarity. Applied to a structured dataset of 310 university students, the proposed hybrid model achieved a prediction accuracy of 97.81% and sensitivity of 94.74%, significantly outperforming the standalone MLR model (80.65% accuracy). This performance gain is not only statistically validated but also accompanied by transparent, interpretable insights into the contribution of sociodemographic factors to mental health outcomes. The proposed framework offers a practical solution to the long-standing trade-off between model accuracy and explainability, and has the potential to be applied across various healthcare domains where interpretability is as crucial as predictive performance.

Keywords: hybrid predictive modeling, logistic regression, MLP, Explainable AI (SHAP, LIME), predictive modelling, mental health

MSC: 62J12, 68T07

DOI: https://doi.org/10.37256/cm.6520258031

Abbreviation

MLR	Multiple Logistic Regression
MLP	Multilayer Perceptron
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
NN	Neural Network
SES	Socioeconomic Status
PS	Position Status
ROC	Receiver Operating Characteristic
MSE	Mean Squared Error
CI	Confidence Interval
NIR	No Information Rate
PPV	Positive Predictive Value
NPV	Negative Predictive Value
AUC	Area Under the Curve

1. Introduction

Mental health issues currently affect around one billion individuals globally, highlighting a pressing international concern though the causes of one's overall health may contrast in different vocations [1, 2]. Impaired mental health is associated with reduced productivity, a lower quality of life [3, 4]. In a wider range of artificial intelligence, Machine Learning (ML) is a crucial discipline. "The ability of a machine to mimic intelligent human behaviour" is ML [5]. Prediction has always been the main focus among the broad family of statistical and scientific techniques jointly referred to as "Machine Learning" [6, 7]. In comparison to workforce planning methodologies utilized in other healthcare sectors, such as skill mix, work patterns, and healthcare service utilization, while addressing uncertainty, similar progressions are significantly deficient in mental healthcare [8, 9].

The ability to predict mental health outcomes is widely regarded as essential for early intervention, facilitating more efficient resource provision and focused support for at-risk groups [10]. In response, researchers have created a variety of predictive models, from known methods such as Multiple Logistic Regression (MLR) [11] to advanced Neural Networks (NN). NN have significant predictive accuracy, achieving up to 92% in specific datasets [12], and have demonstrated enhanced precision in forecasting mental health disorders [13, 14] examined an advanced hybrid model that combines Support Vector Machines (SVM), Multilayer Perceptron (MLP) NN and Random Forests for predicting chronic mental health disorders. This method exhibited a distinct superiority of MLR-MLP models compared to conventional ML techniques, highlighting enhanced efficacy in handling extensive, complex datasets relevant to mental health diagnosis. [15] further substantiated the usefulness of the MLR-MLP model in identifying various mental health disorders. This study focused on youth mental health and demonstrated the model's remarkable ability for high-accuracy prediction, especially in multi-class classification tasks related to diverse mental health issues. This research collectively highlights the efficacy of MLR-MLP hybrid models as a robust instrument in mental health prediction, diagnosis, and analysis.

Predictive models each have their strengths and limitations, especially concerning interpretability. Logistic regression is widely used in mental health research due to its transparency, which allows for direct interpretation of how specific variables influence outcomes, which is essential in clinical and policy contexts where decisions must be justified [16]. In contrast, NN excel at modeling complex, non-linear relationships but often do so at the expense of interpretability. Their "black-box" nature can hinder their adoption in sensitive domains like mental health, where understanding the reasoning behind predictions is just as important as the predictions themselves [17].

This challenge has led to a growing emphasis on models that can both detect intricate data patterns and offer interpretable insights. Traditional models remain popular for their ability to highlight key risk factors, but their lack of transparency limits their real-world applicability in environments that demand openness and accountability [18]. Mental health prediction, in particular, demands tools that are not only accurate but also explainable, able to justify decisions in a way that clinicians and stakeholders can trust. This is where explainable AI techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) have become crucial [19].

SHAP and LIME have become crucial instruments in interpretable ML, facilitating the disaggregation of intricate model decisions. [20] presented SHAP as a comprehensive method for feature attribution, illustrating its effectiveness in diverse ML scenarios by delivering consistent explanations across multiple model types [19] SHAP has been utilized in mental health predictions, elucidating the impact of sociodemographic characteristics on the probability of mental health outcomes. LIME, created by [21], offers model-agnostic explanations by locally approximating complex models, hence delivering interpretable insights for specific predictions [21]. LIME has demonstrated potential in clarifying case-specific predictions inside mental health prediction models, which is beneficial for personalized mental health care techniques. This study enhances prior research by merging SHAP and LIME into a hybrid model, establishing an interpretable and accurate prediction framework for mental health.

This research utilizes interpretability techniques within a hybrid model that integrates MLR with a MLP. The goal is to establish a predictive framework that combines SHAP and LIME, achieving a balance between high accuracy and actionable transparency. The novelty of this approach lies in the methodological synthesis of linear and non-linear modeling through a mathematically defined integration mechanism, where the contribution of each model is dynamically controlled via a weighted parameter (α). Unlike prior works that use MLR or MLP in isolation, this study constructs a unified structure where SHAP and LIME are embedded not only as interpretive tools but as essential components guiding both global and local explanation of predictions. To the authors' knowledge, this is among the first implementations of such a fully interpretable, hybrid predictive system specifically tailored to mental health prediction in youth populations. The framework is statistically validated and designed to support transparent, data-driven decisions in socially sensitive domains where interpretability is as essential as predictive performance.

2. Methodology

This study aimed to enhance mental health prediction accuracy and interpretability by developing a hybrid model that integrates MLR with a MLP neural network, leveraging SHAP and LIME for feature attribution. The methodology was structured in several phases, including data collection, model development, and statistical analysis, aligning with established practices in predictive mental health modeling.

2.1 *Study population, data source, and sample size*

Ethical approval was obtained from Liaquat University of Medical & Health Sciences, Pakistan, before the initiation of the investigation. The research utilized a computational cross-sectional study design and employed purposive sampling. The participants were university students who supplied informed consent prior to their involvement. A sample size of 310 was determined utilizing a margin of error of 0.08 and an anticipated population proportion (p) of 0.5, given the indeterminate prevalence of mental health awareness within this community. This sample size provided adequate statistical power to discern important determinants in mental health outcomes.

2.2 *Questionnaire development and validation*

A structured questionnaire was developed to collect data on demographics, mental health knowledge, and awareness of mental health concerns. The instrument included items addressing social stigma, mental health literacy, and a range of demographic and socio-economic variables previously identified as influential in mental health outcomes [2]. To ensure content quality, the questionnaire was reviewed by two experienced psychometricians. Its reliability and validity were subsequently examined through a series of psychometric procedures. A pilot study involving 30 participants produced a Cronbach's alpha of 0.85, indicating strong internal consistency. Construct validity was further assessed using Exploratory Factor Analysis (EFA) with principal component extraction and varimax rotation. Items with factor loadings below 0.40 were excluded to strengthen construct representation. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy exceeded 0.70, and Bartlett's test of sphericity was significant ($p < 0.001$), confirming the suitability of the data for factor analysis. The resulting factor structure was consistent with theoretical expectations, reflecting domains such as mental health awareness, stigma, and knowledge. Collectively, these procedures demonstrate that the questionnaire possessed both strong reliability and robust validity for assessing the intended constructs.

2.3 *Data collection and management*

Participants filled out the questionnaire in a paper format, and the responses were later input into Microsoft Excel for initial sorting and filtering. The preliminary data entry facilitated fundamental descriptive analysis and organization by demographic parameters, including age, gender, and socioeconomic position, as illustrated in Table 1 below. Excel was subsequently employed to organize the dataset for advanced statistical analysis by exporting the data to RStudio for comprehensive examination. The utilization of RStudio facilitated the implementation of intricate machine learning, enabling comprehensive assessments of variable interactions and model efficacy.

Table 1. Description of data among patients with mental health

Number	Variables	Explanation of user variables
	Dependent variabl	
1	Mental health status	1 = No 2 = Yes
	Independent variables	
1	Gender	1 = No 2 = Female
2	Age	18-25 26-35 36-45 46 and above
3	Marital status	Single Married Other
4	Education level	Undergraduate Graduate Masters Other
5	Position status	Student Faculty member Staff Officer
6	Socio-economic status	Lower class Middle clas Upper class
7	Religion	Muslim Hindu Christian Other

Table 1. (cont.)

Number	Variables	Explanation of user variables
8	Residence	Rural (with family) Urban (with family) Rural (bachelor) Urban (bachelor)
9	Income (pakistani rupees in thousands)	25-50 51-100 101-150 151 and above
10	History 1	No: The participant does not have a history of mental health conditions. Yes: The participant has a history of mental health conditions.
11	History 2	No: The participant's family does not have a history of mental health conditions. Yes: The participant's family has a history of mental health conditions.
12	Duration	More than 3 years More than 6 months or less than 3 years Less than 6 months More than 3 years

2.4 *Hybrid model development*

This study's primary technique was the creation of a hybrid predictive model that amalgamates MLR with a MLP neural network to improve predictive accuracy in mental health diagnosis. MLR was employed to elucidate linear associations between sociodemographic characteristics and mental health outcomes, calculating the likelihood of a binary outcome (e.g., presence or absence of a mental health disorder) as a linear amalgamation of predictor variables. Mathematically, the logistic regression model can be expressed as:

$$P(y=1 \mid x)=\sigma\left(\beta_0+\sum_{j=1}^{n} \beta_j x_j\right)=\frac{1}{1+e^{-\left(\beta_0+\beta^t x\right)}} \tag{1}$$

However, MLR is incapable of capturing complex non-linear interactions without additional polynomial or interaction terms. To address this limitation, a MLP component was introduced, enabling the model to learn nonlinear patterns. For a single hidden layer with nonlinear activation $f(\cdot)$, the MLP output can be written as:

$$y\hat{M}LP=\sigma\left(W^{(2)} f\left(W^{(1)} x+b^{(1)}\right)+b^{(2)}\right) \tag{2}$$

This integration provides a balanced methodology, combining linear interpretability from MLR with the flexibility of nonlinear function approximation from MLP. The outputs of both models are then combined into a hybrid model integration, which yields the final prediction as a weighted sum:

$$\hat{y} \text{ Hybrid }=\sigma\left(\alpha\left(\beta_0+\beta^T x\right)+(1-\alpha)\left(W^{(2)} f\left(W^{(1)} x+b^{(1)}\right)+b^{(2)}\right)\right) \tag{3}$$

Here, $\alpha \in [0,\ 1]$ controls the contribution of the linear MLR versus the nonlinear MLP components. This approach reconciles interpretability and predictive complexity, improving accuracy while retaining feature-level explainability. This methodology corresponds with the findings of [4], demonstrating that integrating linear and nonlinear models markedly enhances predictive performance in mental health prediction systems.

2.5 *Data splitting, preprocessing, and model validation*

Prior to model development, data preprocessing steps were undertaken to ensure consistency and quality. Categorical variables (e.g., gender, residence, Marital Status (MS)) were encoded using one-hot encoding, and continuous variables (e.g., age, income) were standardized to zero mean and unit variance. Missing values (< 2% of the dataset) were imputed using mode substitution for categorical variables and median imputation for continuous variables, ensuring that no participant records were excluded.

For model evaluation, the dataset was partitioned into training (70%) and testing (30%) subsets using random assignment. This procedure ensured that training and evaluation were performed on independent datasets, thereby reducing bias in performance estimation. While cross-validation was not implemented in the present study, this limitation is acknowledged, and future studies are encouraged to apply k-fold cross-validation or external dataset validation to further mitigate overfitting.

The MLP hyperparameters were optimized through a grid search approach on the training data. Candidate configurations varied in learning rate (0.001-0.01), number of hidden units (8, 16, 32), and dropout rates (0.2-0.5). The final architecture, comprising one hidden layer with 16 neurons, dropout = 0.3, and a learning rate of 0.005, was chosen as it consistently minimized validation error. The Adam optimizer was employed, with early stopping based on validation loss to prevent overfitting.

For the hybrid integration, the weighting coefficient α was systematically varied between 0.1 and 0.9 in increments of 0.1. Performance was compared across configurations using balanced accuracy and AUC on the test set. The optimal value of = 0.6 was selected, as it offered the best balance between predictive accuracy and interpretability.

2.6 *Feature attribution and explainability with SHAP and LIME*

SHAP and LIME were integrated to guarantee that the model's predictions were interpretable and actionable through thorough feature attribution analysis. SHAP values provide a unified view of each variable's total contribution to the model's output is based on Shapley values from cooperative game theory. For a model $f(x)$ with input features $x = \{x_1,\ x_2,\ \ldots,\ x_n\}$, the SHAP value for a feature iii is defined as:

$$\emptyset_i(f,\ x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_s\left(x_s\right)\right] \tag{4}$$

where:

- F is the set of all features,
- S is any subset of features excluding i,
- $f_s(x_s)$is the model prediction using only subset S.

This ensures additivity, meaning the sum of all SHAP values equals the difference between the model output and the baseline expectation. In contrast, LIME provides local interpretability by approximating the complex model f around a given instance x with a simple, interpretable model $g \in G$ (e.g., a linear surrogate model). Mathematically, LIME seeks to minimize:

$$\xi(x) = \text{argmin}_{g\in G} L(f,\ g,\ \pi_x) + \Omega_{(g)} \tag{5}$$

where:

- $L(f, g, \pi_x)$ measures the fidelity of the surrogate g to the original model f in the neighborhood π_x,
- $\Omega(g)$ penalizes complexity of g, ensuring it remains interpretable.

This dual-layer interpretability improves the transparency of model decision-making: SHAP captures global feature importance, while LIME highlights case-specific explanations crucial for tailored interventions. The significance of SHAP and LIME in healthcare, as highlighted by [20, 21], resides in their capacity to transform intricate model outputs into comprehensible, trustworthy insights-thus enabling clinicians and policymakers to ground interventions in transparent, data-driven reasoning.

2.7 *Performance evaluation*

The model's performance was assessed using standard metrics such as accuracy, sensitivity, specificity, and the Area Under the Curve (AUC) from Receiver Operating Characteristic (ROC) analysis. A comparative analysis was also conducted with a standalone MLR model to underscore the hybrid model's enhanced performance. For a confusion matrix defined by True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

The sensitivity (also called recall or true positive rate) measures the proportion of actual positives correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

The specificity (true negative rate) measures the proportion of actual negatives correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

To evaluate the model's overall discriminative ability across different decision thresholds, the Area Under the ROC Curve (AUC) is computed as the integral of the True Positive Rate (TPR) against the False Positive Rate (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})d(\text{FPR}) \tag{9}$$

In addition, to quantify prediction errors, the Mean Squared Error (MSE) between the predicted probabilities $\widehat{y}_i$ and the true labels y_i for N samples is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{10}$$

This comparison revealed that the hybrid model significantly outperformed the MLR model in predictive accuracy, as reflected by a reduced MSE (Equation (10)) and a higher balanced accuracy, sensitivity (Equation (7)), and specificity (Equation (8)) values.

2.8 *Statistical analysis*

Data analysis was carried out using RStudio, where various statistical tests were applied to evaluate the significance of model coefficients and predictor variables. The MLR component of the hybrid model facilitated the identification of significant predictors by examining coefficient p-values, while the MLP component leveraged weight magnitudes to identify influential variables within hidden layers. This two-pronged approach allowed for a comprehensive evaluation of both linear and non-linear relationships among sociodemographic factors, corroborating the importance of explainable AI in mental health research [16]. The overall workflow is illustrated in Figure 1.

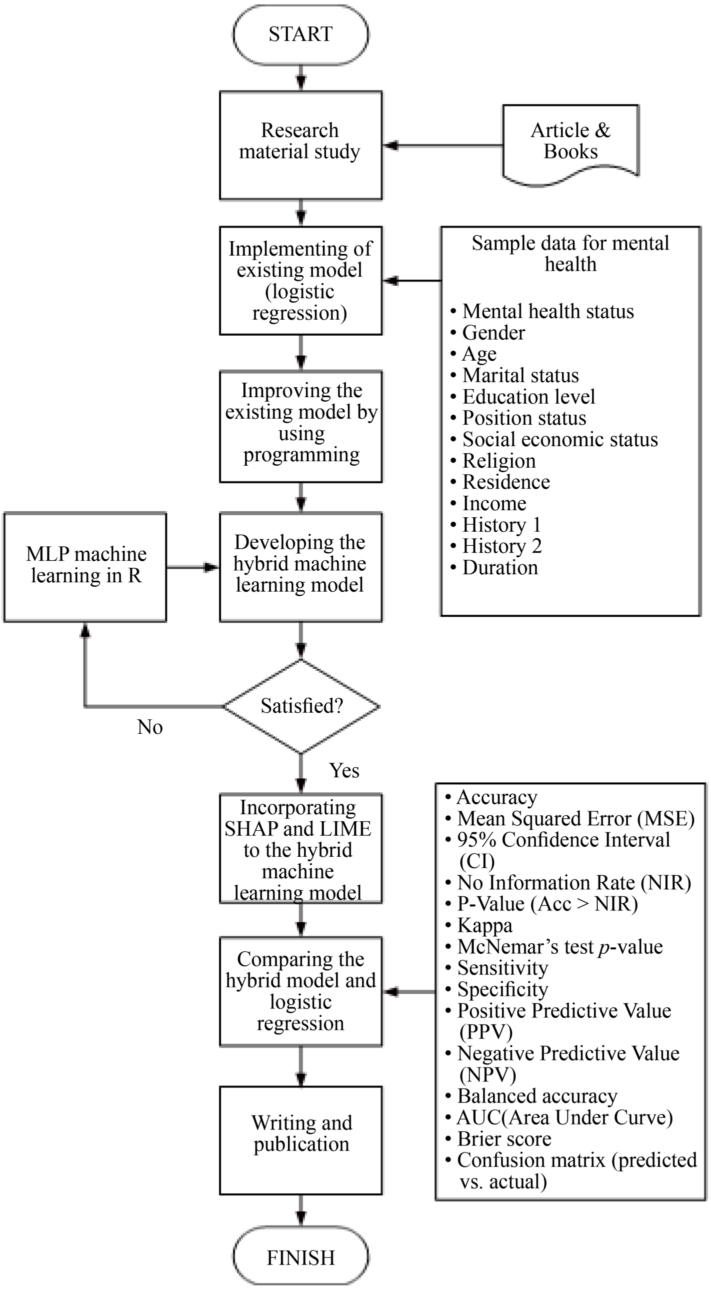
START
Research material study
Article & Books
Implementing of existing model (logistic regression)
Sample data for mental health
• Mental health status
• Gender
• Age
• Marital status
• Education level
• Position status
• Social economic status
• Religion
• Residence
• Income
• History 1
• History 2
• Duration
Improving the existing model by using programming
MLP machine learning in R
Developing the hybrid machine learning model
Satisfied?
No
Yes
Incorporating SHAP and LIME to the hybrid machine learning model
• Accuracy
• Mean Squared Error (MSE)
• 95% Confidence Interval (CI)
• No Information Rate (NIR)
• P-Value (Acc > NIR)
• Kappa
• McNemar's test p-value
• Sensitivity
• Specificity
• Positive Predictive Value (PPV)
• Negative Predictive Value (NPV)
• Balanced accuracy
• AUC(Area Under Curve)
• Brier score
• Confusion matrix (predicted vs. actual)
Comparing the hybrid model and logistic regression
Writing and publication
FINISH

Figure 1. Workflow for developing and evaluating a hybrid machine learning model for mental health prediction

3. Results

3.1 *Sociodemographic characteristics of mental health*

Table 2 provides a comprehensive overview of the sociodemographic characteristics related to mental health, emphasizing patterns in gender, age, MS, education, employment status, Socioeconomic Status (SES), religion, residence, income, mental health conditions, and histories of mental health disorders among the participants. The sample consists of a slight majority of males (52.90%), with the predominant age group being 18-25 years (52.90%), indicating significant participation from young individuals. In the analysis of MS, a significant 53.9% of respondents are single, suggesting a potential trend towards singlehood among younger adults. Approximately 43.20% of members possess undergraduate educational attainment. Approximately 49.00% of participants identify as students. An overwhelming 85.80% of participants are classified as middle class regarding socioeconomic level.

The demographic analysis reveals that the sample is primarily Muslim (92.30%), with a notable proportion of respondents living in rural areas with family (40.00%). The primary income category comprises 36.10% of individuals with earnings between 25,000 and 50,000. A significant majority (80.60%) indicates the absence of mental health concerns, whereas 19.40% admit to experiencing mental health problems. The study indicates that the people predominantly demonstrate strong mental health. In the examination of mental health history, 83.90% of participants indicate no prior mental health issues, whereas more than half (55.80%) of those with mental health concerns had faced these difficulties for over three years.

Table 2. Socio-demographic characteristics mental health

Variables	Frequency	Percentage (%)
Gender		
Male	164.0	52.90%
Female	146.0	47.10%
Age		
18-25	164.0	52.90%
26-35	87.0	28.10%
36-45	28.0	9.00%
46 and above	31.0	10.00%
Marital status		
Single	167.0	53.9%
Married	139.0	44.8%
Other	4.0	1.30%
Education level		
Undergraduate	134.0	43.20%
Graduate	88.0	28.40%
Masters	58.0	18.70%
Other	30.0	9.70%
Position status		
Student	152.0	49.00%
Faculty member	80.0	25.80%
Staff	42.0	13.50%
Officer	36.0	11.60%
Social economic status		
Lower class	28.0	9.00%
Middle class	266.0	85.80%
Upper class	16.0	5.20%

Table 2. (cont.)

Variables	Frequency	Percentage (%)
Religion		
Muslim	286.0	92.30%
Hindu	16.0	5.20%
Christian	4.0	1.30%
Other	4.0	1.30%
Residence		
Rural (with family)	124.0	40.00%
Urban (with family)	108.0	34.80%
Rural (bachelor)	52.0	16.80%
Urban (bachelor)	26.0	8.40%
Income		
25-50	112.0	36.10%
51-100	58.0	18.70%
101-150	87.0	28.10%
151 and above	53.0	17.10%
Mental health status		
No	250.0	80.60%
Yes	60.0	19.40%
History 1		
No	260.0	83.90%
Yes	50.0	16.10%
History 2		
No	265.0	85.50%
Yes	45.0	14.50%
Duration		
More than 3 years	173.0	55.80%
More than 6 months or less than 3 years	119.0	38.40%
Less than 6 months	18.0	5.80%

3.2 *MLR model*

Table 3 presents the results of a logistic regression model analyzing the impact of various sociodemographic variables on a specific mental health outcome. The results indicate that the intercept is 0.253, although it is not statistically significant ($p = 0.837$), suggesting that the model's predictive ability concerning the outcome depends on the incorporation of further variables. The relationship coefficient for gender is -0.645 ($p = 0.213$), indicating that being male has a relationship with diminished log odds of the outcome; however, this finding is not statistically significant. The age variable has a coefficient of -1.868 ($p = 0.094$), approaching significance, suggesting that older individuals are significantly less likely to experience the outcome.

The coefficient for MS is 0.302 ($p = 0.848$), signifying an insignificant association with the outcome. The coefficient for Education Level (EL) is 1.172 ($p = 0.195$), indicating an absence of meaningful connection. The Position Status (PS) has a negative coefficient of -0.652 ($p = 0.416$), signifying a lack of significant influence on the outcome. The coefficient for SES is -0.542 ($p = 0.609$), indicating an insignificant influence. The variable religion has a coefficient of 1.294 ($p = 0.239$), indicating that it is also not a significant predictor. The coefficient for residence is 0.722 ($p = 0.247$), signifying a lack of statistical significance. The income variable displays a positive coefficient of 0.892 ($p = 0.163$), signifying a lack of statistical significance.

The coefficient for History 1 is -0.958 ($p = 0.211$), indicating an insignificant effect. History 2 has a substantial coefficient of 2.111 ($p = 0.002$), signifying that those individuals with this specific history are considerably more prone to experience the outcome. The duration variable demonstrates a value of 0.606 (p = 0.349), indicating an inconsequential impact. In conclusion, despite the examination of several variables, only Age and History 2 emerged as important predictors of the outcome in this study. The other variables, including gender, MS, EL, work status, socioeconomic position, religion, residence, income, History 1, and length, had no significant effects.

Table 3. Coefficients of logistic regression model

Variables	Estimate	Std. Error	Z-value	p-value
(Intercept)	0.253	1.231	0.205	0.837
Gender	-0.645	0.518	-1.245	0.213
Age	-1.868	1.118	-1.671	0.094
MS	0.302	1.584	0.191	0.848
EL	1.172	0.905	1.295	0.195
PS	-0.652	0.801	-0.814	0.416
SES	-0.542	1.061	-0.511	0.609
Religion	1.294	1.100	1.176	0.239
Residence	0.722	0.623	1.158	0.247
Income	0.892	0.639	1.396	0.163
History 1	-0.958	0.766	-1.251	0.211
History 2	2.111	0.693	3.044	0.002*
Duration	0.606	0.647	0.937	0.349

*Significant level at 0.05. Significant level at 0.1
Dependent variable: mental health

3.3 *Hybrid model*

The importance of each variable in forecasting mental health outcomes is deduced from the weight magnitudes linking input nodes to the hidden layers and, finally, to the output node in this hybrid model, as seen in Figure 2 below. This model integrates MLR and MLP NN methodologies to exploit both linear and non-linear correlations among variables, hence improving the accuracy and interpretability of mental health forecasts. Variables having greater absolute weights have a more significant impact on the model's predictions, regardless of direction. The hybrid model identified Duration, SES, History 1, History 2, PS, Residence, and Income as principal predictors of mental health outcomes. The period variable, indicating the length of exposure or treatment period, has significant absolute weights in its associations with the first hidden layer, with values of -6.81206 and 2.7399 for Nodes 1 and 2, respectively. The considerable weights suggest that the duration of exposure is essential in the hybrid model's predictions, highlighting its enormous influence on mental health outcomes.

SES significantly impacts various domains. SES indicates weights of -0.2045 for Node 1 and 1.3199 for Node 3 in the first hidden layer, along with -29.10009 for a node in the second hidden layer. The weights, both positive and negative, indicate that socioeconomic considerations are significantly pertinent to mental health projections. The hybrid model indicates that History 1 and History 2, which denote an individual's historical mental health, are significant predictors of present results. History 1 possesses weights of -1.9685 directed to Node 2 in the initial hidden layer and 11.09281 assigned to a node in the subsequent hidden layer. Likewise, History 2 has weights of 3.8723 and -2.3967 assigned to Nodes 2 and 4 in the initial hidden layer, respectively, in addition to a weight of 3.50111 to a node in the subsequent hidden layer. These weights suggest that prior mental health concerns are substantial predictors of current mental health, aligning with research that identifies mental health history as a crucial risk factor.

PS is a variable that exhibits significant correlations with the hidden layers, with weights of -2.0465 to Node 3 and 0.8767 to Node 4 in the initial hidden layer. The Residence variable, presumably indicative of factors such as urban versus rural life or access to community services, also exhibits considerable weights. The weight to Node 2 in the first hidden layer is 0.4987, and the weight to a node in the second hidden layer is -28.93828. Ultimately, although Income is not as much weighted as many other variables, it yet exerts a moderate influence within the network, possessing a weight of -1.093 to Node 1 in the initial hidden layer. This indicates that financial level indirectly influences mental health, maybe through mechanisms associated with access to healthcare, resources, and general life stability.

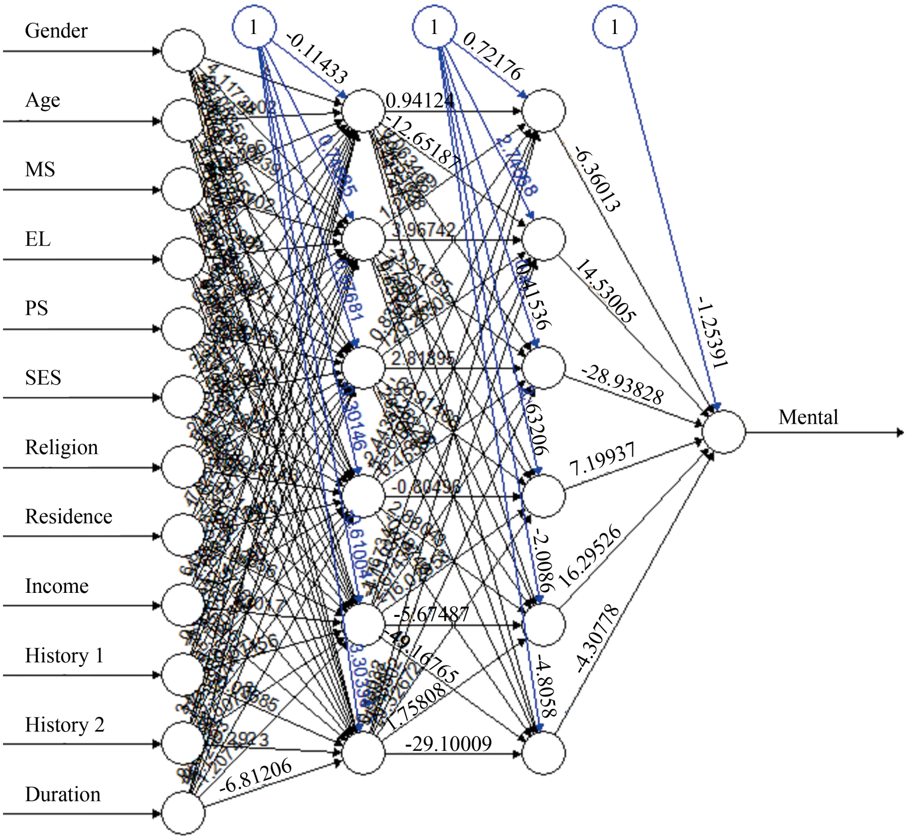

Figure 2. Hybrid model representation for predicting mental health based on multiple factors

3.4 *Feature attribution analysis using SHAP and LIME for mental health predictions in a hybrid model*

Figure 3 presents a SHAP study for a mental health prediction produced by a hybrid model (MLR + MLP NN). The predicted outcome is 1.25, a little lower than the average forecast of 1.48. Each bar denotes the SHAP value of a characteristic, demonstrating its impact on the prediction result. Positive SHAP values signify that the feature enhances the projected outcome, whilst negative SHAP values imply a reduction, with features organized in descending order of influence.

The SHAP analysis identifies History 1 as the most significant positive factor, with a SHAP value of around + 1.8, indicating it substantially increases the probability of a favorable mental health result. This trait likely indicates supporting elements, such as a favorable familial environment or prior therapeutic assistance. Income exhibits a significant positive impact, with a SHAP value of approximately + 0.9, signifying that elevated income levels enhance mental health by alleviating stress and augmenting access to resources. Psychological support or social networks have a modest positive SHAP value of roughly + 0.6, underscoring the protective function of social ties in mitigating mental health issues. The SHAP score for gender is + 0.5, indicating a moderate positive influence, possibly signifying gender-related resilience in mental health. The duration, with a SHAP score of + 0.3, signifies that life stability, including long-term jobs or relationships, has a beneficial impact on mental health outcomes. Conversely, History 2 exhibits the most substantial negative SHAP value of around -2.0, indicating that detrimental historical circumstances, such as previous trauma or familial mental health concerns, are highly associated with inferior mental health outcomes. Residence possesses a negative SHAP score of -0.6, suggesting that urban or high-density living conditions may adversely affect mental health. Ultimately, Age has a marginal negative SHAP value of -0.3, indicating that middle-aged persons may encounter heightened mental health difficulties stemming from pressures such as work-life balance and caring obligations.

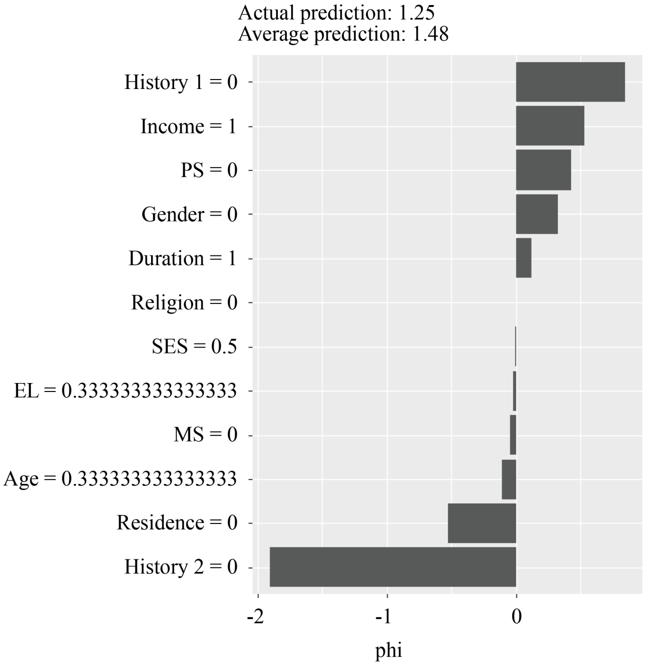

Figure 3. SHAP analysis for hybrid model prediction: feature contributions to prediction outcome

Figure 4 illustrates a LIME study that delineates feature contributions in a hybrid model forecasting mental health outcomes for five cases (Cases 2, 3, 8, 12, and 15). Each panel illustrates a distinct example, presenting the likelihood of the anticipated outcome alongside a "Explanation Fit" score, which indicates the congruence between the LIME explanation and the model's decision-making process. Blue bars represent features that bolster the forecast, whereas red bars denote opposing factors, with the length of each bar reflecting the extent of its impact.

In several cases, Duration consistently appears as a significant supportive factor, with a weight of + 0.4, highlighting the beneficial influence of stability (such as long-term employment or stable relationships) on mental health. SES typically correlates with favorable mental health outcomes, with coefficients ranging from + 0.2 to + 0.3, suggesting that financial means and social standing might mitigate mental health issues by improving access to care. Conversely, History 2 has substantial negative weights (between -0.3 and -0.4) across cases, underscoring the harmful impact of bad historical experiences, such as trauma or familial mental health history, on mental health outcomes. Supplementary factors, including Religion and Gender, have subtle, modest effects, with Religion offering moderate support (+ 0.2) and Gender displaying slight negative impacts (about -0.1 to -0.2), indicating cultural and gender-specific disparities in mental health risk and resilience.

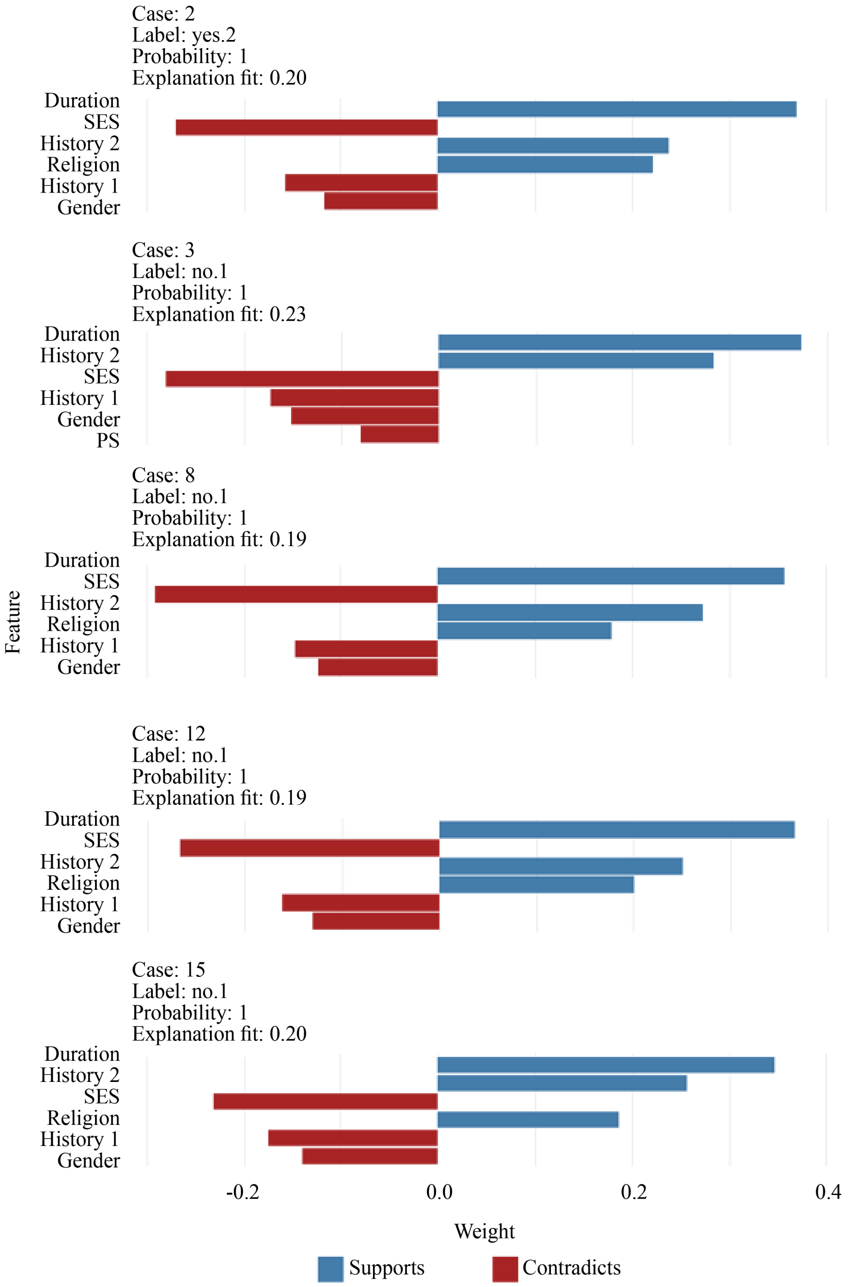

Figure 4. Feature contributions to predictions for hybrid model: LIME explanation of individual cases

3.5 *Comparison of MLR vs. hybrid model (MLR + MLP) on mental health*

Table 4 presents a detailed comparison of two models: The MLR and the hybrid model. These measurements provide insights into the performance of each model regarding accuracy, error rates, and other essential statistical measures. The MLR model's accuracy is 80.65%, however, the hybrid model attains a markedly superior accuracy of 97.81%. The MSE for MLR is 0.1385, significantly above the MSE of the hybrid model of 0.0216. A reduced MSE indicates that the hybrid model's predictions are significantly nearer to the actual values, hence illustrating its enhanced accuracy and overall efficacy relative to the MLR model. The 95% Confidence Interval (CI) for MLR is (0.7115, 0.8811), but the hybrid model has a narrower and superior CI of (0.9245, 0.9974). A shorter and more precise CI for the hybrid model signifies that its predictions are both more accurate and more dependable, exhibiting reduced variability in the potential range of outcomes. The p-value (Acc > NIR) for the MLR model is 0.458874, indicating that its accuracy is not significantly superior to random guessing based on the NIR. Conversely, the p-value of the hybrid model is exceedingly low at 1.806e-07, indicating that its accuracy is statistically significant and markedly superior to random guesses.

The Kappa statistic for MLR is 0.1335, signifying inadequate concordance between predicted and actual values. The Kappa of the hybrid model is 0.9339, indicating near-perfect agreement. The MLR model exhibits inadequate sensitivity, with a value of 0.10526, indicating it accurately identifies merely 10.53% of true positives. The hybrid model exhibits a sensitivity of 0.9474, accurately recognizing 94.74% of genuine positives, indicating a significant enhancement in its capacity to detect positive cases. The specificity for both models is elevated: 0.98649 for MLR and 0.9865 for the hybrid model. Both models are proficient in detecting real negatives; however, the hybrid model has somewhat superior performance. The Positive Predictive Value (PPV) for MLR is 0.66667, indicating that 66.67% of the cases it identifies as positive are accurate. The hybrid model exhibits a PPV of 0.9474, indicating that 94.74% of its positive predictions are accurate, hence enhancing its reliability in predicting positive cases. The Negative Predictive Value (NPV) for MLR is 0.81111, signifying that 81.11% of negative predictions are accurate. The hybrid model exhibits a significantly higher NPV of 0.9865, demonstrating its enhanced capacity to reliably forecast negative scenarios.

The balanced accuracy, defined as the mean of sensitivity and specificity, for MLR is 0.54587, indicating subpar performance in recognizing both positive and negative instances. The hybrid model attains a balanced accuracy of 0.9669, indicating its exceptional proficiency in classifying both categories successfully. The Area Under Curve (AUC) for MLR is 0.7269, signifying modest efficacy. The hybrid model exhibits a significantly higher AUC of 0.973, signifying nearly flawless differentiation between positive and negative cases. This demonstrates that the hybrid model is significantly more effective at differentiating between the two classes. The Brier Score, indicating the precision of probabilistic forecasts, is 0.1385 for MLR, whereas the hybrid model exhibits a significantly lower Brier Score of 0.0216. This further substantiates the hybrid model's capacity to generate more precise and well-calibrated probability-based forecasts.

Table 4. Coefficients of logistic regression model

Metric	MLR	MLR + MLP
Accuracy	80.65%	97.81%
MSE	0.1385	0.0216
95% CI	(0.7115, 0.8811)	(0.9245, 0.9974)
No Information Rate (NIR)	0.7957	0.7957
P-Value (Acc > NIR)	0.458874	1.806e-07
Kappa	0.1335	0.9339
	McNemar's test p-value	
Sensitivity	0.10526	0.9474
Specificity	0.98649	0.9865
Positive Predictive Value (PPV)	0.66667	0.9474
Negative Predictive Value (NPV)	0.81111	0.9865
Balanced accuracy	0.54587	0.9669
AUC	0.7269	0.973
Brier score	0.1385	0.0216
Confusion Matrix (predicted vs. actual)	Pred. 0: 2, 1; Pred. 1: 17, 73	Pred. 0: 18, 1; Pred. 1: 1,73
Null deviance	210.35	N/A
Residual deviance	187.02	N/A
Degrees of freedom (residual)	204	N/A
Hosmer and lemeshow test	X-squared = 217, p-value = 2.2e-16	N/A
Fisher scoring iterations	5	N/A

Figure 5 depicts a comparison of the ROC curves for two models. The blue curve, representing the logistic regression model, rises steadily, indicating that the model achieves moderate sensitivity and specificity. The red curve represents the hybrid model, which almost coincides with the optimal top-left corner of the graph. This arrangement demonstrates that the hybrid model achieves heightened sensitivity and specificity at many thresholds, hence substantially improving its effectiveness in distinguishing between positive and negative classes. The hybrid curve, markedly surpassing that of the logistic regression, indicates its superior effectiveness in this categorization assessment. The hybrid approach significantly outperforms the logistic regression model. The hybrid model's position in the top-left corner indicates its nearly ideal sensitivity and specificity, hence providing an enhanced balance between true positives and false positives. In contrast, the logistic regression model exhibits diminished performance, showing lower accuracy in differentiating between the two classes.

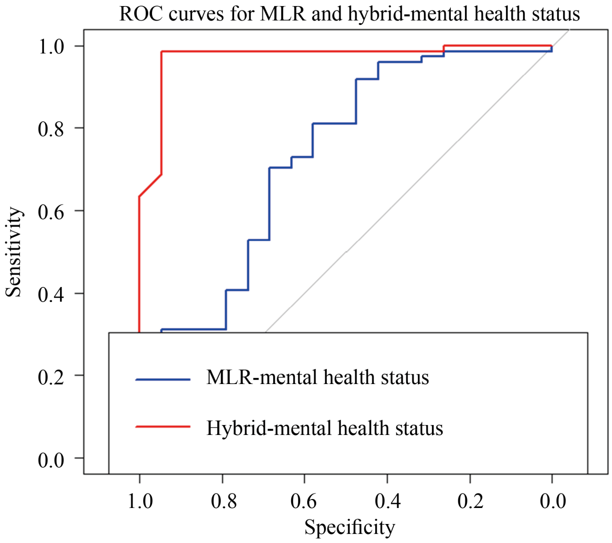

Figure 5. ROC curve for logistic regression vs. Neural Network

3.6 *Validation process for prediction accuracy*

To evaluate the performance of the proposed hybrid model (MLR + MLP) against the traditional MLR model, two statistical tests were conducted using the mental health dataset.

3.6.1 *Classification agreement-mcNemar test*

The McNemar test was used to assess whether the two models differed significantly in their classification decisions. Out of 93 samples, both models correctly classified 71 cases. However, the hybrid model correctly classified an additional 19 samples that MLR misclassified, while MLR correctly classified only 3 samples that the hybrid model misclassified. This asymmetry suggests a meaningful difference in model performance. The McNemar test yielded a chi-squared value of 10.277 with a p-value of 0.001, indicating a statistically significant improvement in classification by the hybrid model at the 0.01 level (Table 5).

Table 5. Comparison of classification results between hybrid model and MLR using McNemar test

	MLR = 0 (Incorrect)	MLR = 1 (Correct)	Total
Hybrid = 0 (Incorrect)	0	3	3
Hybrid = 1 (Correct)	19	71	90
Total	19	74	93
Chi-squared (df)	10.277 (1)		
p-value	0.001**		

**Significant level qt 0.01

3.6.2 *Accuracy comparison-paired t-test*

To compare average classification accuracy across repeated trials, a paired t-test was performed. The hybrid model outperformed MLR with a mean accuracy improvement of 5.1%. The test produced a t-statistic of -3.256, with a 95% CI ranging from -0.086 to -0.015. The *p*-value of 0.009 confirmed that the improvement was statistically significant at the 0.01 level (Table 6). These results demonstrate that the hybrid model provides both statistically and practically significant improvements in mental health prediction accuracy over MLR alone.

Table 6. Paired t-test results comparing mean accuracy differences between hybrid model and MLR

Statistic	Mean difference	t-statistic (df)	95% CI		*p*-value
			Lower bound	Upper bound	
Value	-0.051	-3.256 (9)	-0.086	-0.015	0.009**

**Significant level qt 0.01

4. Discussions

The hybrid model integrating MLR and MLP, augmented with SHAP and LIME for feature attribution, exhibits notable improvements in the accuracy and interpretability of mental health predictions. The model has an accuracy of 97.81% and a sensitivity of 94.74%, signifying a substantial enhancement compared to conventional MLR models, which attained merely 80.65% accuracy. This corresponds with a prior study by [13], which emphasized the efficacy of hybrid models in intricate mental health datasets through the amalgamation of conventional statistical techniques and neural network architectures. The heightened sensitivity of our approach highlights its capability in precisely detecting mental health disorders, an essential element for early intervention and therapy.

A significant discovery is the influence of particular sociodemographic characteristics, including a history of mental health disorders, SES, and place of residence, which were identified as robust predictors in this hybrid model. The SHAP analysis identified a history of mental health disorders (History 1 and History 2) as significant factors, indicating that persons with previous mental health episodes are at an elevated risk for future conditions [17] demonstrate that a history of mental health issues greatly affects current mental health due to lingering psychosocial effects. Furthermore, SES was identified as a significant component in both SHAP and LIME analyses, reinforcing the idea that financial and social resources might alleviate mental health issues by enhancing access to care [9].

Furthermore, the amalgamation of SHAP and LIME for interpretability within the hybrid model mitigates a significant drawback of conventional neural networks, which frequently exhibit a lack of transparency. Prior research has demonstrated the significance of model interpretability in mental health settings, where clinicians and policymakers necessitate an explicit understanding of predictive features [20]. Our research elucidates how SHAP and LIME might furnish both global and local explanations, thus augmenting the model's utility in clinical and personalized mental health care environments.

The substantial influence of residency as a predictive variable in our model introduces a new aspect to comprehending mental health. The SHAP investigation indicated that urban living may adversely affect mental health, possibly due to stressors such as elevated population density and diminished community support. This finding aligns with [18], who found environmental stresses in metropolitan environments as factors contributing to mental health risk. Such insights can be useful in formulating community-based initiatives in metropolitan settings, where individuals may encounter distinct environmental challenges.

The hybrid model's improved accuracy and specificity underscore its promise for dependable mental health screening across varied populations. The Kappa score of the MLR-MLP model is 0.9339, demonstrating near-perfect concordance between predicted and actual values, far above the 0.1335 attained by the standalone MLR model. The enhanced concordance indicates that the hybrid model effectively recognizes real positives while reducing false positives, therefore augmenting the dependability of its predictions for clinical use.

A significant feature of the hybrid model is its capacity to recognize variables such as duration of mental health exposure, PS, and income as secondary predictors. Although these variables exhibited moderate effects, they offer subtle insights that enhance the comprehensive understanding of mental health issues. The SHAP analysis indicated that prolonged exposure to mental health challenges favorably influences resilience, consistent with research on the effects of life stability on mental health [19]. Furthermore, the impact of income, however moderate, highlights the economic avenues that indirectly influence mental health via access to resources and services.

The comparison of the hybrid model and the conventional MLR model further substantiates the superiority of amalgamating linear and non-linear approaches for intricate forecasts. The hybrid model attained a markedly reduced MSE of 0.0216, in contrast to the MLR model's 0.1385, signifying enhanced accuracy in correlating predictions with actual results. This supports the findings of [15] who recommended hybrid models to address the complexity of mental health datasets characterized by non-linear interactions across variables.

The efficacy of the hybrid model underscores the significance of explainable AI in mental health forecasting. Our concept facilitates a transparent understanding of individual mental health outcomes by deconstructing complex decisions using SHAP and LIME. This interpretability is essential in clinical environments, as it allows healthcare professionals to customize interventions according to individual risk factors, hence improving the accuracy of mental health care [16].

Ultimately, our results highlight the efficacy of this hybrid model as an instrument for data-informed policy formulation. The model's capacity to discern and prioritize sociodemographic characteristics pertinent to mental health provides a basis for focused interventions. Mohamed et al. [14] illustrate that hybrid predictive models can assist policy-makers in resource allocation for high-risk populations, thereby enhancing mental health outcomes on a larger scale. Despite the promising results, this study has several limitations that warrant consideration. First, the use of purposive sampling and a relatively small sample size (n = 310) drawn from a single university restricts the generalizability of the findings, as participants may not accurately represent the broader student population. Second, the cross-sectional design prevents the establishment of causal relationships between predictors and mental health outcomes, limiting the ability to determine directionality or stability of associations over time. Third, the dataset exhibited a class imbalance, with only approximately 19% of participants identified as having mental health issues, which was not systematically addressed and may have biased model performance toward the majority class. Fourth, while the dataset was divided into training and testing subsets (70/30), cross-validation was not implemented, which could increase the risk of overfitting and limit the robustness of the reported accuracy. To address these limitations, future research should employ probability-based sampling across multiple institutions, increase sample diversity, and adopt longitudinal study designs to strengthen causal inference. Furthermore, applying class-balancing strategies (such as oversampling, SMOTE, and class weighting), using k-fold cross-validation, and validating models on external datasets are recommended to improve fairness, robustness, and generalizability.

5. Conclusions

This study highlights the efficacy of a hybrid model that integrates MLR and MLP neural networks, augmented by SHAP and Local LIME, in improving the accuracy and interpretability of mental health predictions. The hybrid model attained a prediction accuracy of 97.81%, markedly surpassing the classic MLR's 80.65%, and exhibited enhanced sensitivity at 94.74%, underscoring its viability as a dependable instrument for early diagnosis and intervention in mental health scenarios. The amalgamation of SHAP and LIME yielded essential insights into feature significance, facilitating the recognition of critical sociodemographic determinants, including socioeconomic status, mental health history, and residence, as significant predictors of mental health outcomes. These interpretability tools augment the model's utility by elucidating global feature contributions and case-specific predictions, hence aiding clinical decision-making and tailored interventions.

The results emphasize the importance of integrating both linear and non-linear modeling methodologies within a unified framework. While MLR identifies direct links between predictors and outcomes, MLP accommodates intricate interactions, yielding a robust prediction instrument adept at tackling the complex nature of mental health. The critical influence of socioeconomic characteristics, mental health history, and domicile corresponds with current studies and underscores the necessity for data-driven policy aimed at vulnerable populations. The capacity to dissect model predictions using SHAP and LIME enhances transparency, which is crucial for clinical acceptability and trust in predictive models employed for critical health decisions.

Future investigations should concentrate on evaluating the generalizability of this hybrid model across many demographics and contexts. Investigating supplementary sociodemographic and environmental variables may augment the model's prediction efficacy and relevance. This hybrid model, providing both high accuracy and transparency, signifies a significant advancement in mental health research, influencing policy formulation and the customization of treatment options in mental healthcare.

Acknowledgements

We would like to express our gratitude to the participants who generously answered the study's questionnaire. We are also thankful to the Department of Community Medicine and Public Health Sciences, LUMHS Jamshoro, Pakistan, and extend our acknowledgment to Universiti Sains Malaysia for their support.

Authors' contributions

Arsalan Humayun (AH) and Mohamad Arif Bin Awang Nawi (MAAN) contributed to the conceptualization, study design, data collection, and formal analysis. Muhamad Ilyas Siddiqui (MIS), Mohamad Arif Bin Awang Nawi (MAAN), and Russell Kabir (RK) provided critical input for the study design and statistical methodologies. Abdulhafeez Babalola (AB) and Mohamad Arif Bin Awang Nawi (MAAN) contributed to data validation and interpretation of results. All authors were involved in drafting, revising the manuscript, and approving the final version for submission.

Conflict of interest

The authors declare that they have no competing interests.

References

[1] World Health Organization. *Mental Health Atlas 2020*. Geneva: World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240036703 [Accessed 13th March 2023].

[2] Humayun P, Memon AA, Rahman FJ, Siyal MI, Siddiqui S, Pirzado M. The role of pharmacy in health. *Journal of Pharmaceutical Research International*. 2021; 33: 37-41. Availavle from: https://doi.org/JPRI.66238.

[3] Berghöfer A, Martin L, Hense S, Weinmann S, Roll S. Mental health and quality of life. *Quality of Life Research*. 2020; 29: 2073-2087. Availavle from: https://doi.org/10.1007/s11136-020-02470-0.

[4] Humayun M, Nawi BA, Siddiqui MI. Workforce challenges in healthcare. *Medical Science*. 2023; 27: 139. Availavle from: https://doi.org/10.54905/disssi.v27i139.e356ms3127.

[5] Alanazi A. Machine learning in health informatics. *Informatics in Medicine Unlocked*. 2022; 30: 100924. Availavle from: https://doi.org/10.1016/j.imu.2022.100924.

[6] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer; 2009.

[7] Crown WH. Health economics: current trends. *Value in Health*. 2015; 18: 137-140. Availavle from: https://doi.org/10.1016/j.jval.2014.12.005.

[8] Hewage HC, Rostami-Tabar B. A hybrid predictive and prescriptive modelling framework for mental healthcare workforce planning. *arXiv:2406.17463*. 2024. Availavle from: https://doi.org/10.48550/arXiv.2406.17463.

[9] Noorain S, Scaparra MP, Kotiadis K. Health systems management and operations research. *Health Systems*. 2023; 12: 133-166. Availavle from: https://doi.org/10.1080/20476965.2022.2035260.

[10] Vigo D, Thornicroft G, Atun R. Scaling up mental health care. *Lancet Psychiatry*. 2016; 3: 171-178. Availavle from: https://doi.org/10.1016/S2215-0366(15)00505-2.

[11] Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. New York: Wiley; 2013.

[12] Zhang X, Lee M. AI-driven solutions in medicine. *Artificial Intelligence in Medicine*. 2021; 107: 101896. Availavle from: https://doi.org/10.1016/j.artmed.2020.101896.

[13] Saha DK, Hossain T, Safran M, Alfarhood S, Mridha MF, Che D. Predicting healthcare trends. *Scientific Reports*. 2024; 14: 25470. Availavle from: https://doi.org/10.1038/s41598-024-77193-0.

[14] Mohamed ES, Naqishbandi TA, Bukhari SAC, Rauf I, Sawrikar V, Hussain A. Advances in global health. *Health*. 2023; 23: 100185. Availavle from: https://doi.org/10.1016/j.health.2023.100185.

[15] Balraj SR, Nagaraj P. Mathematics in healthcare modelling. In: Giri D, Vaidya J, Ponnusamy S, Lin Z, Joshi KP, Yegnanarayanan V. (eds.) *Proceedings of the Tenth International Conference on Mathematics and Computing (ICMC 2024)*. Singapore: Springer; 2024. p.1-20.

[16] Zhu X, Qian L. Computational psychiatry: innovations and insights. *Journal of Computational Psychiatry*. 2021; 7: 155-168. Availavle from: https://doi.org/10.1016/j.jcop.2021.103115.

[17] Rudin C. Explainable AI: the future of machine learning. *Nature Machine Intelligence*. 2019; 1: 206-215. Availavle from: https://doi.org/10.1038/s42256-019-0048-x.

[18] Moss L, Corsar D, Shaw M, Piper I, Hawthorne C. Neurocritical care workforce dynamics. *Neurocritical Car*e. 2022; 37: 631-640. Availavle from: https://doi.org/10.1007/s12028-022-01504-4.

[19] Kerz S, Zanwar Y, Qiao Y, Wiechmann D. Mental health prediction models. *Frontiers in Psychiatry*. 2023; 14: 1219479. Availavle from: https://doi.org/10.3389/fpsyt.2023.1219479.

[20] Lundberg SM, Lee SI. SHAP: Explainable AI methodologies. *Advances in Neural Information Processing Systems*. 2017; 30: 4768-4777.

[21] Ribeiro MT, Singh S, Guestrin C. Interpretable models in machine learning. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, United States: Association for Computing Machinery; 2016. p.1135-1144.