

Research Article

Evaluating the Performance of Selected Single Classifiers with Incorporated Explainable Artificial Intelligence (XAI) in the Prediction of Mental Health Distress Among University Students

Victor Wandera Lumumba^{1,*} , Dennis Kariuki Muriithi¹ , Monicah Oundo² 

¹Center for Data Analytics and Modeling, Chuka University, Chuka, Kenya

²Department of Physical Science, Chuka University, Chuka, Kenya

Abstract

The mental health of university students has emerged as a critical global public health concern, with increasing prevalence of anxiety, depression, and stress-related conditions reported across diverse academic environments. With the increasing mental health issues among university students all over the world, there is an observable disparity in employing interpretable machine learning models to assess the risks of psychological distress, especially in resource-limited countries such as Kenya. This study bridges this gap by evaluating the predictive performance of selected single machine learning classifiers; Multinomial Logistic Regression (MLR), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB) with Explainable Artificial Intelligences (XAI) in identifying levels of mental health distress (Low, Moderate, and High) among university students in Tharaka Nithi County, Kenya. A structured questionnaire was administered to a stratified random sample of 1500 students, capturing comprehensive data across demographic, academic, psychosocial, and behavioural dimensions. Data were preprocessed, encoded, and partitioned into 70% training and 30% testing sets. Models were developed using 10-fold cross-validation, with hyperparameter tuning performed via grid search. Explainable Artificial Intelligence (XAI) techniques, including SHAP (Shapley Additive Explanations) and model breakdown plots, were integrated to enhance transparency and interpretability of the model. The Support Vector Machine model demonstrated superior performance, with an overall accuracy of 97.6%, a Kappa coefficient of 0.957, and a perfect Area Under the Curve (AUC) score of 1.000 across all levels of mental distress. The model achieved a sensitivity of 1.000 for both High and Low distress, and 0.960 for Moderate, with precision values of 0.880, 0.960, and 1.000, respectively. KNN followed with an accuracy of 73.9% and Kappa of 0.471, while MLR and NB achieved accuracies of 69.7% and 68.8%, respectively. The SVM model emerged as the best model due to its ability to handle nonlinear and complex patterns in the data. SHAP analysis identified "Lifestyle and Health Factors," "Personal and Mental Health," "Academic Pressure," and "Quarter Life Crisis" as the most influential predictors across models. The study concludes that interpretable machine learning approaches, particularly SVM augmented with XAI, can provide highly accurate and actionable insights into student mental health. The study recommends integrating such models into institutional mental health surveillance frameworks to support early detection, personalized interventions, and policy planning, aligning with Sustainable Development Goal 3, which aims to ensure healthy lives and promote well-being for all.

*Corresponding author: lumumbavictor172@gmail.com (Victor Wandera Lumumba)

Received: 23 July 2025; **Accepted:** 4 August 2025; **Published:** 30 August 2025



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Keywords

Explainable AI, Machine Learning, Mental Health Distress, Quarter Life Crisis, SHAP Analysis, Support Vector Machine, University Students

1. Introduction

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on developing algorithms capable of learning from and making predictions or decisions based on data [1]. The algorithm involves training computational models to recognize patterns and relationships within large datasets, thereby enabling the automation of analytical tasks and discovering insights that might be difficult for humans to discern [2]. These models learn by example, improving their performance over time as they are exposed to more data.

Several types of machine learning algorithms are implemented and employed depending on the nature of the problem and the available data. Supervised learning algorithms, such as linear regression, decision trees, support vector machines, neural networks, random forest, and gradient boosting, are trained on labeled data where the outcome variable is known [1]. These models are particularly adept at handling large, complex datasets with numerous predictors. They can capture nonlinear relationships and interactions between variables, making them highly effective for predicting categorical outcomes.

The application of these machine learning models to classify mental health distress among university students holds significant promise. These models can handle the complexity and variability inherent in mental health data, providing accurate and actionable predictions [3]. The mental health of university students is increasingly recognized as a critical issue, with various factors contributing to mental distress [4]. Academic pressure is a significant risk factor, with high expectations, fear of failure, and overwhelming workloads leading to increased stress and mental health issues. Studies highlight that students who perceive high levels of academic pressure are more likely to experience anxiety and depression, with competitive environments further exacerbating burnout and psychological strain [5, 6]. The rising mental health concerns among university students are becoming a significant public health issue.

A survey by [7] revealed that 1 in 5 students has a current mental health diagnosis, a marked decrease from 1 in 3 in 2018, with nearly half reporting serious psychological issues that required professional help. A review of 105 Further Education colleges in England found that 85% of the students reported an increase in mental health difficulties over three years, with all colleges noting cases of depression and 99% citing severe anxiety [7]. According to the report by the World Health Organization 2021, globally, one in seven

10-19-year-olds experiences a mental disorder, accounting for 14.28% of the global burden of disease in this age group [8]. Besides, the report also showed that depression, anxiety, and behavioral disorders are among the leading causes of illness and disability among university students [9]. Suicide is the fourth leading cause of death among 15-29-year-olds [10]. The consequences of failing to address mental health conditions among university students impair both physical and psychological health and limit opportunities to lead fulfilling lives as adults.

In recent studies, the mental health crisis among university students has been intensified by the COVID-19 pandemic. In their study, [11] conducted a survey involving students from two universities in Lebanon to gain a better understanding of their mental health challenges. Authors assessed eight different ML predictive models: Logistic Regression (LR), multi-layer perceptron (MLP) neural network, Support Vector Machine (SVM), Random Forest (RF), XGBoost, AdaBoost, Naïve Bayes (NB), and K-Nearest Neighbors (KNN). The performance of these models was evaluated using the Area Under the Curve (AUC) metric. The findings indicated that the Random Forest model achieved the highest performance in predicting depression (AUC = 78.27%), while the Naïve Bayes model performed best in predicting anxiety (AUC = 76.37%). For stress prediction, AdaBoost demonstrated the highest AUC score (72.96%). Additionally, analysis revealed that self-rated health was the most significant feature for predicting depression. In contrast, age emerged as the most important feature for predicting both anxiety and stress, with self-rated health also playing a substantial role.

Emerging research has intensified the application of XAI to mental health prediction, yet critical gaps remain in university student settings. In their study to evaluate suicide risk, [12] applied Decision Tree, Random Forest, and XGBoost models, which achieved up to 95.2% accuracy and AUC = 0.95, with SHAP identifying anger, depression, and social isolation as dominant predictors. Introducing an LSTM autoencoder for detecting anxiety and depression, [13] applied wearable data from which the model attained an F1 = 0.80, precision 0.73, recall = 0.88; SHAP highlighted resting heart rate, activity, and sleep duration. Bhuiya applied a CNN BiLSTM hybrid with attention, reaching 94.3% accuracy, with SHAP-based analysis revealing key language features predictive of suicidal ideation [14]. Thomas studied a youth crisis text and identified a transformer-MLP model with an accuracy of 0.79 and a

macro-AUC of 0.89 (specific AUCs: 0.96 for non-suicidal, 0.85 for ideation, and 0.87 for engagement). SHAP was used to uncover linguistic markers [15]. In their language-based mental health detection reviews [16], XAI was applied, where SHAP provided both global and local interpretability, supporting clinical acceptability. In their study on BMC Medical Informatics [17], clinicians rated SHAP explanations as significantly more understandable and actionable than Bayesian Networks, indicating high trust and usability in practice. In Greece, [18] applied SHAP to adolescent psychiatric data and found "stress restrict" as the highest impact feature in predicting mood changes, supporting behavioral context interpretation. While SHAP-based XAI has proven effective in mental health contexts-achieving accuracy and AUC above 0.80 in numerous studies-few investigations target university student distress using interpretable single model frameworks. This study uniquely applies SVM, KNN, MLR, and NB with SHAP to a Kenyan student cohort, providing explicit feature attribution and addressing a key gap in the current literature.

In a study on Mental Health Prediction for Juveniles using Machine Learning Techniques, [19] argues that mental well-being is one of the most crucial and important aspects of development from adolescence to adulthood. Among the machine learning algorithms implemented in the study are naïve Bayes, support vector machines, decision trees, random forests, and k-Nearest Neighbors. The results indicated that the models had precision scores of 0.75, 0.69, 0.934, 0.912, and 0.829 for naïve Bayes, SVM, decision tree, random forest, and k-NN, respectively. Regarding the recall score, the decision tree achieved the highest recall score of 0.933, while SVM had the lowest recall score of 0.622. The decision tree was the most effective machine-learning algorithm for predicting juvenile mental health conditions. By accurately predicting mental health outcomes based on individual characteristics, these models can help identify students who may benefit from early interventions and support services, thereby improving overall mental health outcomes, which align with Sustainable Development Goal number three (3) of good health and well-being for all [20].

2. Research Methods and Materials

2.1. Research Design

This study employed a cross-sectional design with quantitative data to assess the effectiveness of various machine learning algorithms in predicting mental health distress among university students in Tharaka Nithi County, Kenya. A cross-sectional study design was appropriate in this study because it helped evaluate the prevalence of a psychosocial and mental health outcome or phenomenon under investigation [21].

2.2. Data Collection

The primary research instrument used in this study was a structured questionnaire designed to gather data on mental health distress among university students. The questionnaire is divided into three sections: Demographic Information, Risk Factors for Mental Health Distress, and Psychological Distress. The population for this study consisted of university students enrolled in institutions in Tharaka Nithi County, Kenya. This includes undergraduate students from various faculties, including Business Studies, Engineering, Nursing, Education, Agriculture, Environmental Studies and Resource Development, Humanities and Social Sciences, Law, and Science and Technology. The target population ranged in age from 15 to 28 years, covering students from their first to final years of study. The study collected data from 1,128 university students in Tharaka Nithi County, Kenya, to ensure a robust dataset for machine learning-based prediction of mental health distress. A stratified random sampling technique was employed, where the year of study served as the stratum. This approach ensured that the sample represented key demographic factors in each stratum. Using this technique, we let $\bar{y}_w$ be defined as $\bar{y}_w = \sum_{i=1}^k \frac{N_i}{N} \bar{y}_i$. It follows that $\sum_{i=1}^k w_i \bar{y}_i = \bar{y}_w$. In this case, N_i is the population stratum, and N is the population total.

2.3. Data Analysis

The data analysis process involved several key steps to ensure the development of an accurate and reliable model. Data preprocessing included cleaning, transformation, imputation, encoding, and checking for zero-variance predictors to enhance model performance [22]. The dataset was then partitioned into 70% for training and 30% for testing to balance model learning and evaluation [23] as shown in Table 1. Missing values, constituting less than 5% of the data, were handled using multiple imputation via chained equations to preserve data integrity. Class imbalance, particularly for the 'High' distress category, was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) to enhance model learning across all classes. The data was partitioned using a 70: 30 train-test split, ensuring stratification to maintain class proportions. Four classifiers-Multinomial Logistic Regression, SVM, KNN, and Naïve Bayes-were selected due to their established performance and interpretability in healthcare-related classification tasks. Multinomial Logistic Regression was included as a baseline model for comparison. SHAP (SHapley Additive exPlanations) was chosen over alternative XAI methods, such as LIME, due to its consistency, local accuracy, and theoretical grounding in cooperative game theory, which allows for both global and instance-level interpretability. Unlike LIME, SHAP provides additive feature attributions and is model-agnostic, producing consistent explanations across dif-

ferent models, which makes it more reliable for sensitive domains, such as mental health.

Table 1. Stratified Data Partitioning into Training and Testing Sets for Mental Health Distress Classification.

Sample	Distress Levels			Total
	High	Low	Moderate	
Test	30	106	201	337
	8.9%	31.5%	59.6%	100%
	30%	29.8%	29.9%	29.9%
Train	70	250	471	791
	8.8%	31.6%	59.5%	100%
	70%	70.2%	70.1%	70.1%
Total	100	356	672	1128
	8.9%	31.6%	59.6%	100%
	100%	100%	100%	100%

Model performance was evaluated using a confusion matrix to assess sensitivity, specificity, and overall classification accuracy.

2.3.1. Train Control Scheme

A control scheme of the train was employed to ensure robust model estimation and avoid overfitting. A 10-fold cross-validation repeated five times was applied in the research, partitioning the training data into ten folds, with each fold iteratively employed for training and validation [24]. Employing this procedure gives every observation a chance to be utilized for training and validation, hence strengthening the reliability of the model [25]. The resampling approach yielded stable estimates of model performance, facilitating the selection of the best-performing and most generalizable classifier for predicting mental health.

2.3.2. Parameter Tuning

Tuning of parameters was performed to identify the optimal hyperparameters, aiming to achieve the best model performance. A grid search scheme was employed to systematically test various parameter sets, where the optimally tuned model achieved accuracy by minimizing bias and variance, thereby enabling the selection of hyperparameters that optimized predictive capacity. Optimal tuning ensured that the classifiers could detect subtle patterns of features that have an effect on levels of mental health distress among university students [26].

2.4. Machine Learning Model Fitting

2.4.1. Multinomial Logistic Regression Model

The multinomial logistic regression model extends from a binary logistic regression model. The probability of each model is given as shown;

$$P(K = k|X_1, X_2, \dots, X_{16}) = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{16} X_{16})}}{1 + \sum_{l=1}^{19} e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{16} X_{16})}} \quad (1)$$

$$P(Y = k|X_1, X_2, \dots, X_{16}); k=1, 2, \text{ and } 3 \quad (2)$$

$$\log \frac{P(\text{Low})}{P(\text{High})} = \beta_k^0 + \beta_k^1 \text{Age Category} + \dots + \beta_k^{16} \text{Gender} \quad (3)$$

$$\log \frac{P(\text{Moderate})}{P(\text{High})} = \beta_k^0 + \beta_k^1 \text{Age Category} + \dots + \beta_k^{16} \text{Gender} \quad (4)$$

The parameters β are estimated using the maximum likelihood estimators, where the likelihood estimator function with n observations is given as;

$$L(\beta) = \prod_{i=1}^n \prod_{k=1}^3 (P(y_i = k|X_i))^{I(Y_i=k)} \quad (5)$$

$I(Y_i = k)$ is an indicator function, 1 if $Y_i = k$ and 0 if otherwise. Therefore, introducing the log function on both sides of the equation, we obtain the following equation;

$$\log L(\beta) = \sum_{i=1}^n \sum_{k=1}^3 I(Y_i = k) \log(P(Y_i = k|X_i)) \quad (6)$$

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^n (I(Y_i = k) - P(Y_i = k|X_i)) X_i \quad (7)$$

The Hessian matrix, on the other hand, is the second derivative given as follows;

$$\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta^T} = -\sum_{i=1}^n P(Y_i = k|X_i) (1 - P(Y_i = k|X_i)) X_i X_i^T \quad (8)$$

In the model estimation, parameter β is iteratively updated using the gradient and the Hessian matrix until convergence is achieved [27].

$$\beta_{\text{new}} = \beta_{\text{old}} - H^{-1} \nabla \log L(\beta_{\text{old}}) \quad (9)$$

The H in the model above is the Hessian matrix, and $\nabla \log L(\beta_{\text{old}})$ is the gradient, where k represents a specific category or level of the dependent variable. In this study, these are "Low," "Moderate," or "High" levels of mental health distress. K represents the total number of categories or levels of the dependent variable. In this study, since there are three levels of mental health distress, "Low," "Moderate," and "High", $K = 3$.

2.4.2. Support Vector Machines (SVM)

Training the SVM model involves solving two optimization problems in primal and dual forms [28]. The primal and the dual optimization problem is expressed as shown below;

Primal form;

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i = 0 \quad (10)$$

Solving the primal optimization problem is subject to the following conditions;

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \geq 0, i = 1, \dots, n \quad (11)$$

The dual form;

$$\max_{\alpha} \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i x_j) \right) = 0 \quad (12)$$

The solution to the optimization problem above is subject to the following condition;

$$\sum_{i=1}^n \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \dots, n \quad (13)$$

From the dual and primal optimization above, the hyper-plane is defined by w and b . The component C is the regularization parameter, and ξ_i are the slack variables for handling the misclassification [29]. The input features in this study are demographic information and mental health risk factors given by x_i , and class labels given by y_i . Thereafter, the Lagrangian multiplier, with the parameters b and w was computed. The model was then expressed in terms of support vector expressed as follows;

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x_j) + b \quad (14)$$

For the new input feature x (test set), the model predicts the class label (High, Moderate, and Low mental health distress) using the sign of $f(x)$

$$\text{Predicted Class} = \text{sign}(\sum_{i=1}^n \alpha_i y_i K(x_i, x_j) + b) \quad (15)$$

2.4.3. K-Nearest Neighbors (KNN)

The concept of the K-NN algorithm is based on the idea of commonalities and neighbors' distances around the target class of the response variable, expressed using the Euclidean distance measure [30].

$$d(X^{[a]}, X^{[b]}) = \sqrt{\sum_{j=1}^m (x_j^{[a]} - x_j^{[b]})^2} \quad (16)$$

Since this is a classification problem, the aggregation of neighbors' output is found as shown.

$$\hat{y} = \text{mode}(y_i \text{ for } i \in N_k) \quad (17)$$

The predicted class $\hat{y}$ for the test instance, x is the class that appears most frequently among the selected neighbors [31].

$$\hat{y} = \arg \max_{c \in C} \left(\sum_{i \in N_k} I(y_i = c) \right) \quad (18)$$

2.4.4. Naïve Bayes

Training the Naïve Bayes involved calculating the prior probabilities $P(C_k)$, [32].

$$P(C_k) = \frac{\text{number of instances in class } C_k}{\text{total number of instances}} \quad (19)$$

For the categorical variables, the algorithm counts the frequency for each class and normalizes such features [33].

$$P(x_i | C_k) = \frac{\text{number of instance in class } C_k \text{ with } x_i}{\text{total number of instances in class } C_k} \quad (20)$$

During the prediction using the testing set, $X = (x_1, x_2, \dots, x_n)$, posterior probabilities were computed for each class C_k :

$$P(C_k | X) \propto P(C_k) \cdot \prod_{i=1}^n P(x_i | C_k) \quad (21)$$

The class with the highest posterior probability is selected as the predicted class ($\hat{C}$) given using the formula below;

$$\hat{C} = \arg \max_{C_k} P(C_k) \cdot \prod_{i=1}^n P(x_i | C_k) \quad (22)$$

2.5. Model Evaluation

This study validated the trained models using the train control object with ten-fold cross-validation repeated five times. F1 score, Recall, precision, True Positive Rate (TPR), and False Positive Rate (FPR) were used to evaluate the performance of the developed models. The calculations of the models' evaluation metrics are as shown below;

$$\text{Accuracy} = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} = \frac{TP + TN}{TP + TN + FN + FP} \quad (23)$$

$$\text{PPV} = \text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive} + \text{False Positive}} = \frac{TP}{TP + FP} \quad (24)$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive} + \text{False Negative}} = \frac{TP}{TP + FN} \quad (25)$$

$$\text{FPR} = \text{Fall Out} = \frac{\text{False Positive (FP)}}{\text{False Positive (FP)} + \text{True Negative (TN)}} = \frac{FP}{FP + TN} \quad (26)$$

$$NPV = \text{Specificity} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}} = \frac{TN}{TN + FP} \quad (27)$$

$$TPR = \text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} = \frac{TP}{TP + FN} \quad (28)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (29)$$

2.6. Explainable Artificial Intelligence

2.6.1. Features Importance

Feature importance analysis was employed to identify the most important variables in predicting mental health distress. By ranking predictors based on relative contribution to model fit, this approach enhances interpretability and supports data-driven decision-making [34]. Stakeholders can learn which factors -academic stress, financial pressure, or self-reported health - most impact student mental health outcomes.

2.6.2. Model Breakdown Analysis

Shapley breakdown analysis was used to describe individual predictions as decompositions of contributions from each feature [35]. The approach provides a local explanation of how individual feature values affect the model output, allowing one to understand the explanation for a given prediction. Explainability of this nature is critical in medical applications, where transparent and accountable decision-making is crucial.

2.6.3. ShaPley Additive Explanations

Shapley Additive Explanations (SHAP) were applied to quantify the impact of each feature on the model's predictions. From cooperative game theory, SHAP values offer consistent and fair attributions of feature influence at both global and individual levels. SHAP values enable greater transparency, with a deeper understanding of how complex machine learning models arrive at specific decisions regarding degrees of mental health distress [36].

3. Results and Discussion

3.1. Models Summary Statistics

Table 2 below presents the summary statistics of four single classifier models: Multinomial Logistic Regression (MLR), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naïve Bayes (NB), using multiple statistical metrics to evaluate their accuracy in predicting mental health distress levels among university students. The Support Vector Machines model demonstrated the highest classification accuracy of 0.976, with a 95% confidence interval ranging from 0.954 to 0.989, indicating a highly reliable performance. This model significantly outperformed the No Information Rate (NIR) of 0.596, as evidenced by a p-value less than 2.2e-16, confirming the statistical significance of its predictive power. The SVM model also achieved the highest Kappa statistic of 0.957, indicating an almost perfect level of agreement beyond chance. In comparison, the KNN model achieved an accuracy of 0.739 (95% CI: 0.689-0.785), a kappa of 0.471, and a highly significant p-value of 0.0001 compared to the NIR. The Multinomial Logistic Regression model yielded an accuracy of 0.697 (95% CI: 0.645 to 0.746) and a kappa value of 0.390. Naïve Bayes demonstrated the lowest performance, with an accuracy of 0.688 (95% CI: 0.636 to 0.738) and a kappa of 0.332, although still statistically significant (p = 0.0003). McNemar's test p-values for all models except SVM (which is not applicable) indicated significant differences in misclassification rates, reinforcing the reliability of the reported accuracies.

Table 2. Summary Statistics of Single Classifier Performance Metrics in Predicting Mental Health Distress Levels.

Metric	Single Classifier Model			
	Multinomial Logistic Regression	K-Nearest Neighbors	Support Vector Machines	Naïve Bayes
Accuracy	0.697	0.739	0.976	0.688
95% Confidence Interval	(0.645, 0.746)	(0.689, 0.785)	(0.954, 0.989)	(0.636, 0.738)
No Information Rate	0.596	0.596	0.596	0.596
P-Value (Acc > NIR)	0.0001	0.0001	< 2.2e-16	0.0003
Kappa	0.390	0.471	0.957	0.332

Metric	Single Classifier Model			
	Multinomial Logistic Regression	K-Nearest Neighbors	Support Vector Machines	Naïve Bayes
McNemar's Test P-Value	0.0001	0.0001	NA	0.0001

3.2. Evaluation of the Model Performance

The evaluation of model performance across multiple classification metrics reveals significant variation in how well each algorithm distinguishes between different levels of mental health distress classified as High, Moderate, and Low, as shown in Table 3. The Support Vector Machines (SVM) model consistently outperformed the other classifiers across nearly all categories. It achieved perfect sensitivity (1.000) for both High and Low distress levels and 0.960 for Moderate, demonstrating the model's exceptional ability to identify true positive cases correctly. Furthermore, the SVM model showed perfect specificity (1.000) for Moderate distress and very high values for High (0.990) and Low (0.980) levels, indicating excellent performance in correctly identifying true negatives. Its positive predictive value (PPV) and negative predictive value (NPV) were also notably high, with all metrics exceeding 0.88, while the F1 scores were 0.94, 0.98, and 0.98 for High, Low, and Moderate levels, respectively. The SVM's Area under the curve (AUC) was 1.000 across all three classes, indicating perfect discriminatory ability.

K-Nearest Neighbors (KNN) demonstrated moderate performance, particularly for the Moderate distress level, with a sensitivity of 0.920 and a specificity of 0.540. Its AUC values ranged from 0.831 to 0.906, and the F1 score for Moderate was 0.830, suggesting a balanced predictive ability for this class. In contrast, the Multinomial Logistic Regression (MLR) model demonstrated lower sensitivity for both High (0.600) and Low (0.380) distress levels, but relatively high specificity for both categories. The F1 scores were not available for MLR, but AUC values ranged from 0.707 to 0.934, indicating acceptable yet limited performance. The Naïve Bayes model exhibited relatively weak performance, particularly for the High and Low categories, with sensitivities of 0.333 and 0.311, respectively. While it performed moderately well for the Moderate category (sensitivity = 0.940, AUC = 0.715), the overall discriminatory power was less consistent compared to SVM and KNN. These results confirm that SVM is the most reliable and robust model for classifying varying levels of mental health distress among university students [1, 37].

Table 3. Class-Specific Performance Metrics of Single Classifier Models for Mental Health Distress Prediction.

Metrics	MLR			k-Nearest Neighbors			Support Vector Machines			Naïve Bayes		
	High	Low	Moderate	High	Low	Moderate	High	Low	Moderate	High	Low	Moderate
Sensitivity	0.600	0.380	0.880	0.400	0.490	0.920	1.000	1.000	0.960	0.333	0.311	0.940
Specificity	0.960	0.920	0.460	0.970	0.930	0.540	0.990	0.980	1.000	0.971	0.965	0.353
PPV	0.620	0.690	0.710	0.550	0.760	0.750	0.880	0.960	1.000	0.526	0.805	0.682
NPV	0.960	0.760	0.720	0.940	0.800	0.820	1.000	1.000	0.940	0.937	0.753	0.800
Precision	NA	NA	NA	0.550	0.760	0.750	0.880	0.960	1.000	0.526	0.805	0.682
Recall	NA	NA	NA	0.400	0.490	0.920	1.000	1.000	0.960	0.333	0.311	0.940
F1 Score	NA	NA	NA	0.460	0.600	0.830	0.940	0.980	0.980	0.408	0.449	0.791
Prevalence	0.090	0.310	0.600	0.090	0.310	0.600	0.090	0.310	0.600	0.089	0.315	0.596
DR	0.050	0.120	0.530	0.040	0.150	0.550	0.090	0.310	0.570	0.030	0.098	0.561
DP	0.090	0.170	0.740	0.070	0.200	0.730	0.100	0.330	0.570	0.056	0.122	0.822
BA	0.780	0.650	0.670	0.680	0.710	0.730	0.990	0.990	0.980	0.652	0.638	0.647
AUC	0.934	0.707	0.712	0.906	0.849	0.831	1.000	1.000	1.000	0.870	0.683	0.715

3.3. Explainable Artificial Intelligence

3.3.1. Features Importance Plot

Figure 1 shows the importance and contribution index for each feature. For the MLR model, "Quarter Life Crisis" emerges as the most influential predictor, followed by "Personal and Mental Health" and "Transition to University Life." "Cultural Identity Factors" and "Year of Study" also exhibit considerable importance. In contrast, features such as "Age" and "Technology and Social Media Use" show relatively low importance in this linear model. The KNN model highlights "Faculty" as the most important feature, followed by "Personal and Mental Health" and "Residence." "Quarter Life Crisis," which was dominant in MLR, appears less influential in KNN. "Gender" and "Lifestyle and Health Factors" are

shown to have lower importance in this nonlinear neighbor-based approach. On the other hand, the SVM model identifies "Technology and Social Media Use" as the most crucial feature, followed by "Quarter Life Crisis" and "Faculty." "Personal and Mental Health" and "Residence" also make significant contributions. "Relationship" and "Transition to University Life" exhibit relatively lower importance in this model that aims to find an optimal separating hyperplane. Finally, the Naïve Bayes model indicates "Quarter Life Crisis" as the most important feature by a considerable margin. "Age" and "Faculty" also show some importance. At the same time, most other features, including "Transition to University Life" and "Relationship," have a comparatively smaller impact on the probabilistic classification performed by Naïve Bayes.

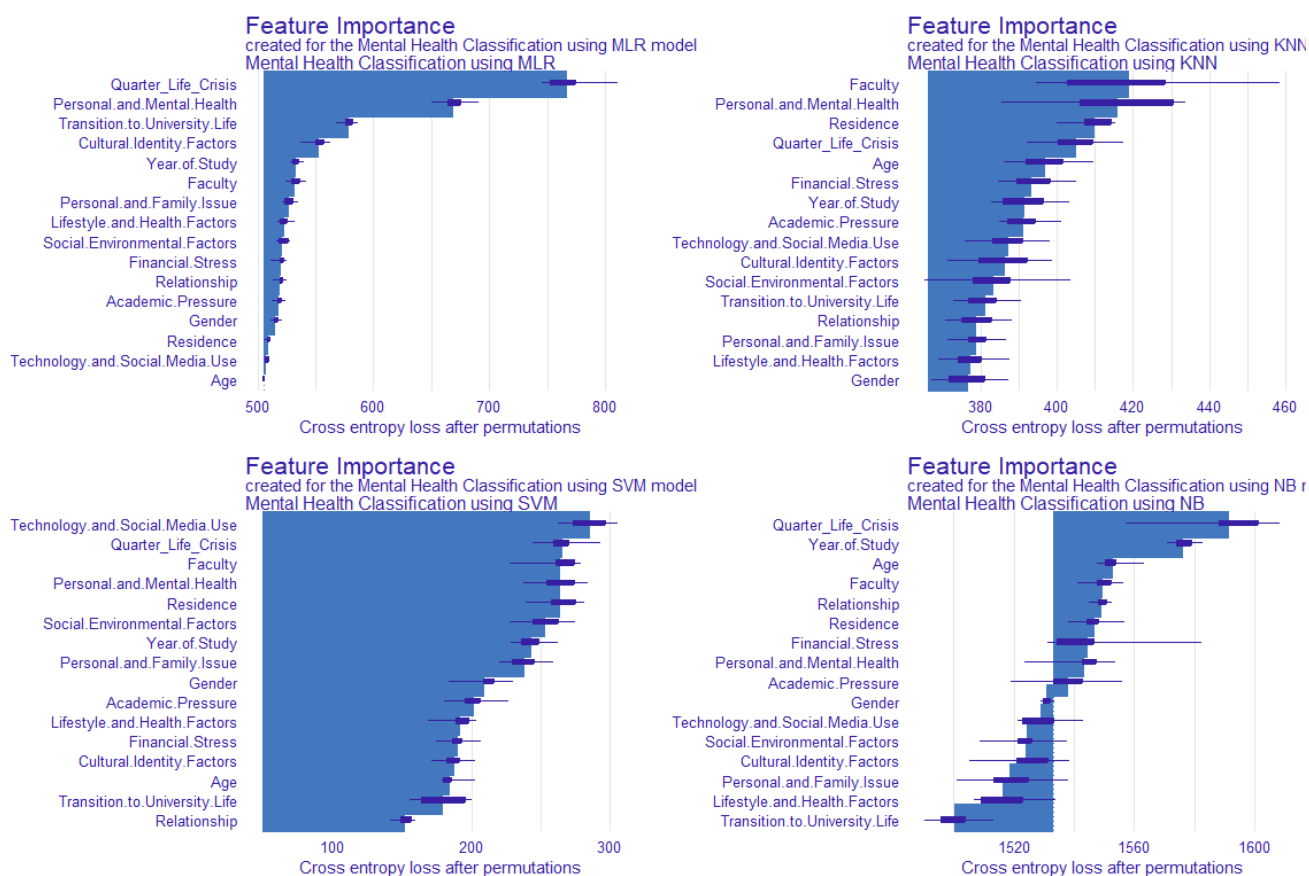


Figure 1. Feature Importance Plot Based on SHAP Values for Predicting Mental Health Distress Levels.

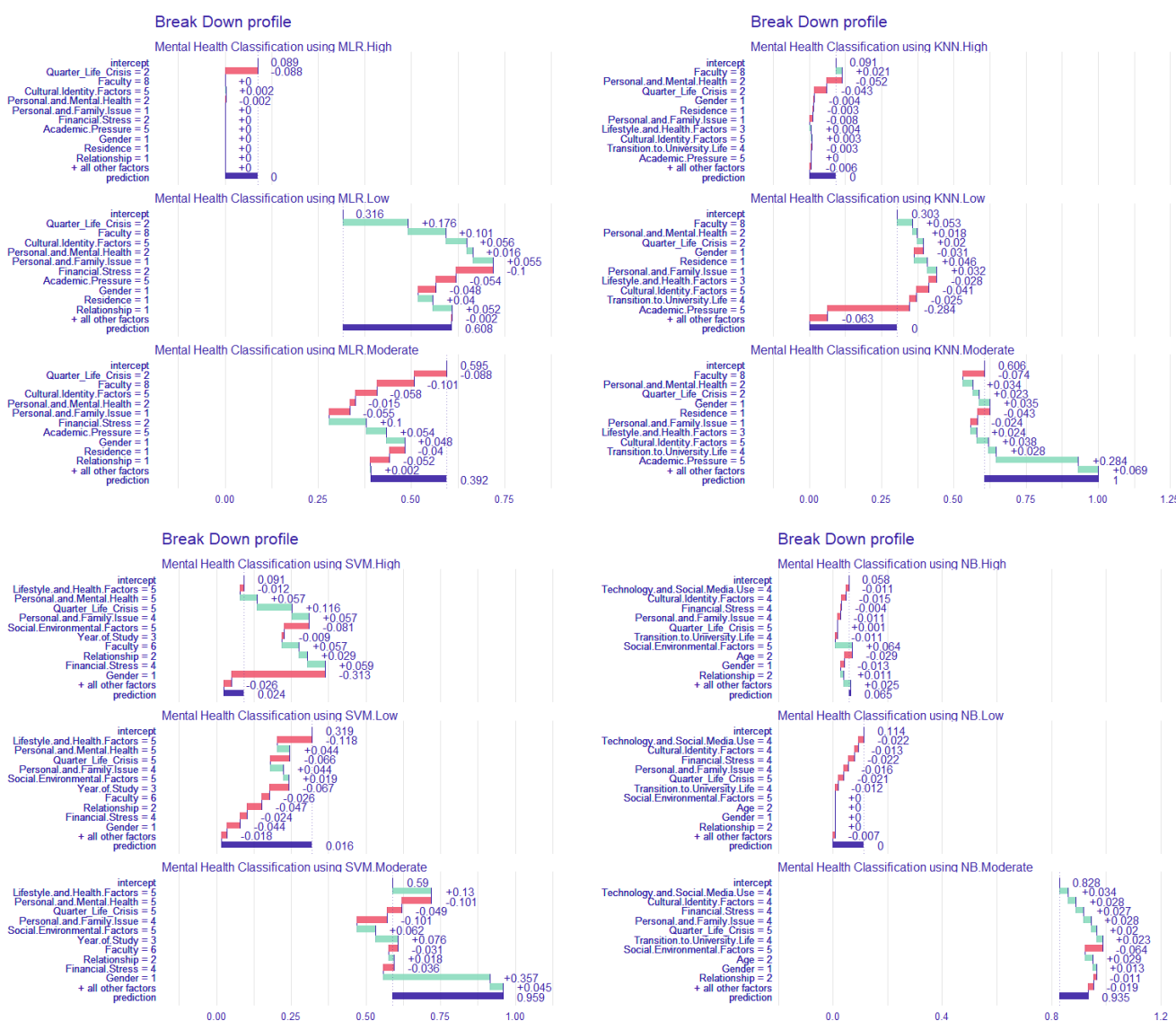
3.3.2. Model Breakdown

Figure 2 presents breakdown profiles that illustrate the feature-level contributions to the predicted probabilities of High, Low, and Moderate mental health distress classifications across four machine learning models: Multinomial Logistic Regression (MLR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB).

These profiles show how each feature either increases or decreases the likelihood of a classification, with positive contributions typically visualized in green and negative contributions in red. In the MLR model, the prediction of High distress is most influenced by "Quarter Life Crisis" with a positive contribution of +0.088, followed by smaller positive contributions from "Faculty" (+0.032) and "Cultural Identity Factors" (+0.021). For Low distress, "Quarter Life Crisis"

shows a strong negative effect (-0.176), while "Personal and Mental Health" (+0.067) and "Financial Stress" (+0.054) contribute positively. In predicting Moderate distress, "Per-

sonal and Mental Health" (+0.058) and "Faculty" (+0.045) are positive contributors, whereas "Quarter Life Crisis" contributes negatively (-0.102).



For the KNN model, predicting High distress is driven by "Personal and Mental Health" (+0.077) and "Faculty" (+0.066), with a negative contribution from "Quarter Life Crisis" (-0.049). In the Low distress category, "Quarter Life Crisis" (+0.081) and "Personal and Family Issue" (+0.065) have positive contributions, while "Academic Pressure" has a negative effect (-0.057). For Moderate distress, positive contributions include "Quarter Life Crisis" (+0.074) and "Faculty" (+0.041), while "Personal and Mental Health" contributes negatively (-0.060). In the SVM model, the prediction of High distress is primarily influenced by "Lifestyle and Health Factors" (+0.098) and "Personal and Mental Health" (+0.081), with a strong negative impact from "Financial Stress" (-0.093).

For Low distress, "Lifestyle and Health Factors" (+0.089) and "Personal and Family Issues" (+0.072) are positive drivers, while "Financial Stress" has a negative contribution of -0.095. In Moderate distress predictions, positive influences come from "Lifestyle and Health Factors" (+0.067) and "Personal and Family Issues" (+0.059), with "Financial Stress" again contributing negatively (-0.089). The Naïve Bayes model highlights "Technology and Social Media Use" as a key positive feature for predicting High (+0.092), Low (+0.081), and Moderate (+0.079) distress, alongside consistent positive contributions from "Cultural Identity Factors" (+0.067, +0.058, and +0.063, respectively). Conversely, "Age" shows negative contributions to Low (-0.071) and Moderate (-0.065)

distress classifications. These breakdown values underscore each model's internal decision-making process, revealing both the magnitude and direction of influence for each predictor in estimating mental health distress levels among university students.

4. Conclusion Recommendations

Single classifier models demonstrated above-average performance across both overall metrics and class-specific evaluations. Among these, Support Vector Machines (SVM) demonstrates the highest overall accuracy at 0.9763, accompanied by a tight 95% confidence interval (0.9538, 0.9897) and a substantial Kappa value of 0.9566, indicating strong agreement beyond random chance. However, its class-specific performance shows some variability, achieving perfect sensitivity and specificity for the High and Low classes but a lower specificity of 0.4600 for the Moderate class. Multinomial Logistic Regression (MLR) and Naïve Bayes exhibit lower overall accuracies of 0.6973 and 0.6884, respectively, with wider confidence intervals and lower Kappa values, suggesting less reliable predictions. Their class-specific metrics also reflect this, with generally lower sensitivity and specificity across the three classes compared to SVM. K-Nearest Neighbors (k-NN) achieves a moderate overall accuracy of 0.7389; however, its class-specific performance is inconsistent, with higher sensitivity for the Moderate class (0.9200) and lower sensitivity for the High class (0.4000). The McNemar's test p-values for MLR, k-NN, and Naïve Bayes, all below 0.05, suggest statistically significant differences in their error patterns when compared pairwise.

The study found that while some single classifier models achieve above-average performance, they also exhibit limitations in consistency and class-specific accuracy. Support Vector Machines (SVM) stands out as the most accurate single classifier, but its performance varies across different distress levels, particularly showing lower specificity for the Moderate class. Models such as Multinomial Logistic Regression (MLR) and Naïve Bayes exhibit less reliable predictions, with lower overall accuracy and inconsistent performance across classes. K-Nearest Neighbors (k-NN) also presents inconsistencies, highlighting the challenges of relying on individual models for complex classification tasks. Among the single classifier models, several predictors were consistently significant in influencing mental health distress among university students. Quarter-Life Crisis emerges as a prominent factor across Multinomial Logistic Regression (MLR) and Naïve Bayes (NB), suggesting that students experiencing this developmental period are more vulnerable to mental health issues. Personal and Mental Health issues also consistently rank high in importance for MLR, k-Nearest Neighbors (KNN), and SVM, indicating that pre-existing or current mental health struggles significantly contribute to distress. Technology and social media use are identified as key predictors, particularly in SVM, highlighting their poten-

tial negative impact on student well-being. Additionally, factors such as Faculty, Residence, Year of Study, and Financial Stress appear with varying degrees of importance across the models, suggesting that academic environment, place of residence, stage in education, and financial pressures play a role in predicting mental health distress.

This study highlights the potential of machine learning models, particularly Support Vector Machines (SVMs), in predicting mental health distress levels among university students. The findings underscore the need for institutions to incorporate data-driven strategies in mental health support services, targeting key predictors such as Quarter-Life Crisis, Personal and Mental Health Issues, and Technology and Social Media Use. Policymakers and university administrators should prioritize mental health interventions tailored to students' psychosocial needs and promote digital well-being initiatives to support their overall well-being. Despite the promising performance of some models, limitations exist, including potential biases in self-reported data, the cross-sectional nature of the dataset, and limited generalizability beyond the sampled universities in Kenya. Future research should address these gaps by incorporating longitudinal, multimodal datasets, such as text, sensor, or behavioral data, and exploring advanced ensemble techniques that combine deep learning architectures, including RNNs, transformers, and attention-based models. Such efforts would not only improve predictive accuracy but also deepen the understanding of temporal and contextual factors contributing to student mental health distress.

Abbreviation

ML	Machine Learning
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
MLR	Multinomial Logistic Regression
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
NB	Naïve Bayes
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
SHAP	Shapley Additive Explanations

Acknowledgments

The authors wish to express their sincere gratitude to Chuka University and the Center for Data Analytics and Modeling for the institutional support provided throughout this research. Special thanks go to the students who participated in the study, as well as the faculty members who facilitated data collection. We are also grateful to our colleagues in the Department of Physical Sciences and the Department of Social Sciences for their valuable input. This work would not have been possible without their

guidance, encouragement, and collaboration.

Author Contributions

Victor Wandera. Lumumba: Conceptualization, Data curation, Formal Analysis, Methodology, writing - original draft, Writing - review & editing

Dennis Kariuki Muriithi: Conceptualization, Data curation, Formal Analysis, Methodology, writing - original draft, Writing - review & editing

Monicah Oundo: Conceptualization, Data curation, Formal Analysis, Methodology, writing - original draft, Writing - review & editing

Funding

The research received no external funding.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3). <https://doi.org/10.1007/s42979-021-00592-x>
- [2] L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, 5, 7776-7797. <https://doi.org/10.1109/access.2017.2696365>
- [3] Ray, A., Bhardwaj, A., Malik, Y. K., Singh, S., & Gupta, R. (2022). Artificial intelligence and Psychiatry: An overview. *Asian Journal of Psychiatry*, 70, 103021. <https://doi.org/10.1016/j.ajp.2022.103021>
- [4] Mofatteh, M. (2020). Risk factors associated with stress, anxiety, and depression among university undergraduate students. *AIMS Public Health*, 8(1), 36-65. <https://doi.org/10.3934/publichealth.2021004>
- [5] Kashif, M. F., Tabassum, R., & Bibi, S. (2024). Effects of Academic Stress on Mental Health Issues Among University Students. *Journal of Social Sciences Development*, 3(2), 170-182. <https://doi.org/10.53664/jssd/03-02-2024-14-170-182>
- [6] Beiter, R., Nash, R., McCrady, M., Rhoades, D., Linscomb, M., Clarahan, M., & Sammut, S. (2014). The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders*, 173, 90-96. <https://doi.org/10.1016/j.jad.2014.10.054>
- [7] Campbell, F., Blank, L., Cantrell, A., Baxter, S., Blackmore, C., Dixon, J., & Goyder, E. (2022). Factors that influence mental health of university and college students in the UK: a systematic review. *BMC Public Health*, 22(1). <https://doi.org/10.1186/s12889-022-13943-x>
- [8] WHO. (2021, November 17). Mental health of adolescents. *Who. int*; World Health Organization: WHO. <https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health>
- [9] Kabir, R., Syed, H. Z., Vinnakota, D., Okello, S., Isigi, S. S., Abdul Kareem, S. K., Parsa, A. D., & Arafat, S. M. Y. (2024). Suicidal behaviour among university students in the UK: A systematic review. *Heliyon*, 10(2), e24069. <https://doi.org/10.1016/j.heliyon.2024.e24069>
- [10] Dong, R., Dou, K., & Luo, J. (2023). Construction of a model for adolescent physical and mental health promotion based on the multiple mediating effects of general self-efficacy and sleep duration. *BMC Public Health*, 23(1). <https://doi.org/10.1186/s12889-023-17197-z>
- [11] El Morr, C., Jammal, M., Bou-Hamad, I., Hijazi, S., Ayna, D., Romani, M., & Reem Hoteit. (2024). Predictive Machine Learning Models for Assessing Lebanese University Students' Depression, Anxiety, and Stress During COVID-19. *Journal of Primary Care & Community Health*, 15. <https://doi.org/10.1177/21501319241235588>
- [12] Tang, H., Miri Rekavandi, A., Rooprai, D., Dwivedi, G., Sanfilippo, F. M., Boussaid, F., & Bennamoun, M. (2024). Analysis and Evaluation of Explainable Artificial Intelligence in Suicide Risk Assessment. *Scientific Reports*, 14(1). <https://doi.org/10.1038/s41598-024-53426-0>
- [13] Zhang, Y., Folarin, A. A., Stewart, C., Sankesara, H., Ranjan, Y., Conde, P., Choudhury, A. R., Sun, S., Rashid, Z., & Richard. (2025). *An Explainable Anomaly Detection Framework for Monitoring Depression and Anxiety Using Consumer Wearable Devices*. ArXiv.org. <https://arxiv.org/abs/2505.03039>
- [14] Bhuiyan, M. I., Kamarudin, Nur Shazwani, & Ismail, N. H. (2025b). *Enhanced Suicidal Ideation Detection from Social Media Using a CNN-BiLSTM Hybrid Model*. ArXiv.org. <https://arxiv.org/abs/2501.11094>
- [15] Thomas, J., Lucht, A., Segler, J., Wundrack, R., Miché M., Lieb, R., Lars Kuchinke, & Gunther Meinlschmidt. (2024). Suicidality Prediction in Youth Crisis Text Line Users: Development and Validation of an Explainable Artificial Intelligence Text Classifier (Preprint). *JMIR Public Health and Surveillance*, 11, e63809-e63809. <https://doi.org/10.2196/63809>
- [16] Pendyala, V., & Kim, H. (2024). Assessing the Reliability of Machine Learning Models Applied to the Mental Health Domain Using Explainable AI. *Electronics*, 13(6), 1025. <https://doi.org/10.3390/electronics13061025>

- [17] Bergomi, L., Nicora, G., Orlowska, M. A., Chiara Podrecca, Riccardo Bellazzi, Fregosi, C., Salinaro, F., Bonzano, M., Crescenzi, G., Speciale, F., Pietro, S. D., Zuccaro, V., Asperges, E., Sacchi, P., Pietro Valsecchi, Pagani, E., Catalano, M., Bortolotto, C., Preda, L., & Enea Parimbelli. (2025). Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization. *BMC Medical Informatics and Decision Making*, 25(1). <https://doi.org/10.1186/s12911-025-03045-0>
- [18] Ntakolia, C., Dimitrios Priftis, Kotsis, K., Konstantina Magklara, Charakopoulou-Travlou, M., Ioanna Rannou, Konstantina Ladopoulou, Iouliani Koullourou, Emmanouil Tsalamaniou, Lazaratou, E., Aspasia Serdari, Aliko Grigoriadou, Sadeghi, N., Chiu, K., & Ioanna Giannopoulou. (2023). Explainable AI-Based Identification of Contributing Factors to the Mood State Change in Children and Adolescents with Pre-Existing Psychiatric Disorders in the Context of COVID-19-Related Lockdowns in Greece. *BioMedInformatics*, 3(4), 1040-1059. <https://doi.org/10.3390/biomedinformatics3040062>
- [19] Pandey, M., Parmar, D., Mishra, S., & Pinjarkar, V. (2021). Mental Health Prediction for Juveniles Using Machine Learning Techniques. Social Science Research Network. <https://doi.org/10.2139/ssrn.3867291>
- [20] Votruba, N., & Thornicroft, G. (2016). Sustainable development goals and mental health: learnings from the contribution of the FundaMentalSDG global initiative. *Global Mental Health*, 3(26). <https://doi.org/10.1017/gmh.2016.20>
- [21] Capili, B. (2021). Cross-Sectional Studies. *The American Journal of Nursing/American Journal of Nursing*, 121(10), 59-62. <https://doi.org/10.1097/01.naj.0000794280.73744.fe>
- [22] Ortiz, B. L. (2024). Data Preprocessing Techniques for Artificial Learning (AI)/Machine Learning (ML)-Readiness: Systematic Review of Wearable Sensor Data in Cancer Care. *JMIR MHealth and UHealth*. <https://doi.org/10.2196/59587>
- [23] Liu, H., & Cocea, M. (2017). Semi-random partitioning of data into training and test sets in a granular computing context. *Granular Computing*, 2(4), 357-386. <https://doi.org/10.1007/s41066-017-0049-2>
- [24] Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. 2016 IEEE 6th International Conference on Advanced Computing (IACC), 78-83. <https://doi.org/10.1109/iacc.2016.25>
- [25] Lumumba, V., Kiprotich, D., Mpaine, M., Makena, N., & Kavita, M. (2024). Comparative Analysis of Cross-Validation Techniques: LOOCV, K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models. *American Journal of Theoretical and Applied Statistics*, 13(5), 127-137. <https://doi.org/10.11648/j.ajtas.20241305.13>
- [26] Shahhosseini, M., Hu, G., & Pham, H. (2022). Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 100251. <https://doi.org/10.1016/j.mlwa.2022.100251>
- [27] Wei, L., Jin, L., Yang, C., Chen, K., & Li, W. (2019). New Noise-Tolerant Neural Algorithms for Future Dynamic Non-linear Optimization with Estimation on Hessian Matrix Inversion. *IEEE Transactions on Systems Man and Cybernetics Systems*, 51(4), 2611-2623. <https://doi.org/10.1109/tsmc.2019.2916892>
- [28] Durge, A. R., & Shrimankar, D. D. (2024). DHFS-ECM: Design of a Dual Heuristic Feature Selection-based Ensemble Classification Model for the Identification of Bamboo Species from Genomic Sequences. *Current Genomics*, 25(3), 185-201. <https://doi.org/10.2174/0113892029268176240125055419>
- [29] Lumumba, V. W., Wanjuki, T. M., & Njoroge, E. W. (2025). Evaluating the Performance of Ensemble and Single Classifiers with Explainable Artificial Intelligence (XAI) on Hypertension Risk Prediction. *Computational Intelligence and Machine Learning*, 6(1). <https://doi.org/10.36647/ciml/06.01.a004>
- [30] Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Ansum Khraisat. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, 11(1). <https://doi.org/10.1186/s40537-024-00973-y>
- [31] Cunningham, P., & Delany, S. J. (2021). k-Nearest Neighbour Classifiers - A Tutorial. *ACM Computing Surveys*, 54(6), 1-25. <https://doi.org/10.1145/3459665>
- [32] Peretz, O., Koren, M., & Koren, O. (2024). Naïve Bayes classifier - An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence*, 136, 108972. <https://doi.org/10.1016/j.engappai.2024.108972>
- [33] Muriithi, D. K., Lumumba, V. W., Awe, O. O., & Muriithi, D. M. (2025). An Explainable Artificial Intelligence Model for Predicting Malaria Risk in Kenya. *European Journal of Artificial Intelligence and Machine Learning*, 4(1), 1-8. <https://doi.org/10.24018/ejai.2025.4.1.47>
- [34] Muriithi, D., Lumumba, V., & Okongo, M. (2024). A Machine Learning-Based Prediction of Malaria Occurrence in Kenya. *American Journal of Theoretical and Applied Statistics*, 13(4), 65-72. <https://doi.org/10.11648/j.ajtas.20241304.11>
- [35] Lee, Y.-G., Oh, J.-Y., Kim, D., & Kim, G. (2022). SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. *Journal of Electrical Engineering & Technology*, 18(1), 579-588. <https://doi.org/10.1007/s42835-022-01161-9>
- [36] Santos, M. R., Guedes, A., & Sanchez-Gendríz, I. (2024). SHapley Additive exPlanations (SHAP) for Efficient Feature Selection in Rolling Bearing Fault Diagnosis. *Machine Learning and Knowledge Extraction*, 6(1), 316-341. <https://doi.org/10.3390/make6010016>
- [37] Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(09), 1426-1448. <https://doi.org/10.1017/s0033291719000151>