Computer Science and Information Technologies
Vol. 6, No. 3, November 2025, pp. 335~345
ISSN: 2722-3221, DOI: 10.11591/csit.v6i3.p335-345

A machine learning approach for early prediction of mental health crises

Hassan Chigagure[1], Lucy Charity Sakala[2]
[1]Department of Computer Science, School of Information Science and Technology, Harare Institute of Technology, Harare, Zimbabwe
[2]Department of Computer Science, Faculty of Science and Engineering, Bindura University of Science Education, Bindura, Zimbabwe

Article Info

Article history:

Received Jun 3, 2025
Revised Jun 30, 2025
Accepted Jul 13, 2025

Keywords:

Ensemble methods
Feature selection
Longitudinal hospital data
Machine learning
Mental health crisis prediction

ABSTRACT

The global mental health crisis, intensified by the COVID-19 pandemic, placed unprecedented strain on healthcare systems and highlighted the urgent need for proactive crisis prevention strategies. This study investigated the effectiveness of various machine learning (ML) models in predicting mental health crises within 28 days post-hospitalization, leveraging an eight-year longitudinal dataset. Multiple data preprocessing techniques, including feature selection (EFSA, RFECV), imputation, and class imbalance handling (SMOTE, Tomek links), were systematically applied to enhance model performance. Six traditional classifiers—logistic regression, support vector machine, k-nearest neighbors, naive Bayes, XGBoost, and AdaBoost—were evaluated alongside ensemble learning (EL) methods (bagging, boosting, stacking). Performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC were used for comprehensive assessment. Results demonstrated that ensemble methods, particularly boosting and bagging, consistently achieved high predictive accuracy (up to 93%), with XGBoost and AdaBoost emerging as top performers. Feature selection and class imbalance techniques further improved model robustness and generalizability. The findings underscored the potential of ML-driven approaches for early identification of at-risk patients, enabling more effective resource allocation and timely interventions in mental health care. Recommendations for integrating these predictive tools into clinical workflows were discussed to support data-driven decision-making.

Corresponding Author:

Lucy Charity Sakala
Department of Computer Science, Faculty of Science and Engineering
Bindura University of Science Education
Bindura, Zimbabwe
Email: lsakala@hit.ac.zw

1. INTRODUCTION

Approximately one billion people globally live with a mental disorder [1]. The COVID-19 pandemic has significantly intensified this worldwide mental health emergency, leading to an escalating need for mental healthcare services. This surge in demand is putting a strain on healthcare systems, which are already grappling with a shortage of qualified mental health personnel [2]. Mental disorders have the potential to significantly impact various aspects of life, encompassing academic or professional achievement, interactions with family and friends as well as engagement within the community [3], [4]. The global economy suffers a loss of $1 trillion annually due to decreased productivity caused by two prevalent mental disorders: anxiety and depression [5]. When considering overall mental issues, including reduced

productivity and the burden of illness, the global economic burden associated with this issue was estimated to be around $2.5 trillion in 2010, a figure projected to climb to $6 trillion by 2030 [6].

Prompt intervention can often prevent the worsening of symptoms that lead to mental health crises and subsequent hospitalization [7]. Unfortunately, patients often only access urgent care, such as a hospital or psychiatric facility, when they are already in the midst of a crisis [8]. At this stage, preventative measures are no longer an option, which limits the ability of psychiatric services to effectively allocate their already strained resources [9], [10]. Therefore, a key step in improving patient outcomes and managing caseloads is to identify individuals at risk of a crisis before it happens [11]. In busy clinical environments, manually reviewing vast amounts of patient data to make proactive care decisions is simply not feasible; it is unsustainable and prone to errors [9], [12]. Shifting these tasks to automated analysis of hospital records, however, offers a promising solution. This approach could revolutionize healthcare by enabling continuous, large-scale data review [13]. Research has already shown we can predict critical health events for many conditions, like hypertension, diabetes, circulatory failure, hospital readmission, and even in-hospital death [9], [14]. However, when it comes to mental health, the existing research mainly focuses on predicting specific events such as suicide, self-harm, or a first episode of psychosis [11]. We do not have as much information on continuously predicting the wider range of mental health crises that need urgent care or hospitalization. A lot is still unknown about whether we can continuously use machine learning to estimate the risk of an impending mental health crisis [15]. If we could, it would allow us to better allocate healthcare staff and potentially prevent crises from even happening [16]. In addition, it is not yet clear if new predictive technologies would be truly useful for mental healthcare practitioners, especially concerning their impact on health outcomes or long-term cost savings [17], [18].

Current clinical practice often relies on retrospective symptom assessment and self-reporting, which can be unreliable and reactive. Identifying individuals at high risk of impending mental health crises (e.g., severe depressive episodes, psychotic breaks, and suicidal ideation) is challenging due to the complex interplay of clinical, behavioural, social, and environmental factors. The lack of predictive tools leads to delayed intervention meaning patients receive care late, often when their condition is severe, requiring more intensive and costly interventions like inpatient hospitalization [9]. Additionally, increased suffering: Individuals endure prolonged periods of distress and functional impairment, and inefficient resource allocation- healthcare resources are often deployed reactively rather than strategically, leading to bottlenecks and potential burnout for mental health professionals.

This study addresses important gaps by conducting a comprehensive comparison of machine learning models using an eight-year longitudinal dataset to predict mental health crises. It aims to identify key risk factors that contribute most to crisis prediction and evaluate the performance of different models. The findings provide valuable insights and practical recommendations for effectively integrating machine learning into mental health care systems. This approach has potential to improve early detection and timely intervention for at-risk patients.

2. METHODOLOGY

2.1. Research design

This study, a retrospective cohort study, focused on developing and assessing mental health crisis prediction models using existing health records. Retrospective cohort studies analyze already collected data to evaluate outcomes based on prior exposures or conditions. This research employs a quantitative methodology, leveraging a machine learning approach to predict mental health crises among patients within a 28-day period following hospitalization [9], [19]. Using this methodology allows for efficient analysis of large dataset and timely identification of at-risk patients, improving predictive accuracy and healthcare intervention.

2.2. Feature selection

Feature selection involves choosing a subset of pertinent features from a larger pool of available features within a dataset [20]. It involves identifying and retaining the most informative and impactful features while discarding or disregarding less relevant or redundant ones [20]–[22]. Feature selection aims to enhance the performance, interpretability, and efficiency of machine learning models by concentrating on the most impactful aspects of the data [23]. This study tested two feature selection methods, and both were used for model construction. The two methods evaluated were the ensemble of feature selection algorithms (EFSA) and recursive feature elimination with cross-validation (RFECV).

2.3. Experimental procedure

The procedure began with the dataset (comprising datasets from eight different years) being loaded into Google Colab. These datasets were then concatenated together to create a unified dataset. The next step involved data preprocessing, a crucial phase that encompasses several essential tasks [24]. Data cleaning was

performed to remove any inconsistencies or outliers, followed by data imputation to fill in missing values. Label encoding was applied to make the data numerical, and data exploration was conducted to unveil underlying trends and hidden patterns within the data [25]. The analysis also assessed class imbalance and involved feature selection to enhance model performance. Subsequently, the experiment moved on to model construction, where a variety of traditional ML models, including LR, SVM, K-NN, NB, XGBoost, and AdaBoost, were implemented. Furthermore, ensemble learning (EL) techniques, such as bagging, boosting, and stacking, were utilised to combine these models [26]. The next phase was model evaluation, in which the model's performance using an array of metrics such as accuracy, precision, recall, F1 score, kappa, geometric mean, and AUC-ROC was assessed [27]. This extensive experimental setup was created to make it easier to assess the study's results, improve their validity, and determine whether they might be replicated in other research settings. Google Colab Research was used to carry out empirical experiments.

Figure 1 illustrates the workflow diagram outlining the entire experimental procedure used in this study. It begins with loading and preprocessing the dataset, followed by feature selection to identify the most relevant predictors. The workflow then proceeds to model development, including hyperparameter tuning and training multiple classifiers. Finally, the models are evaluated using performance metrics, and the results are analysed to determine the best predictive approach. Thus, Figure 1 provides a comprehensive overview of the systematic steps undertaken to ensure robust and reproducible machine learning experiments.

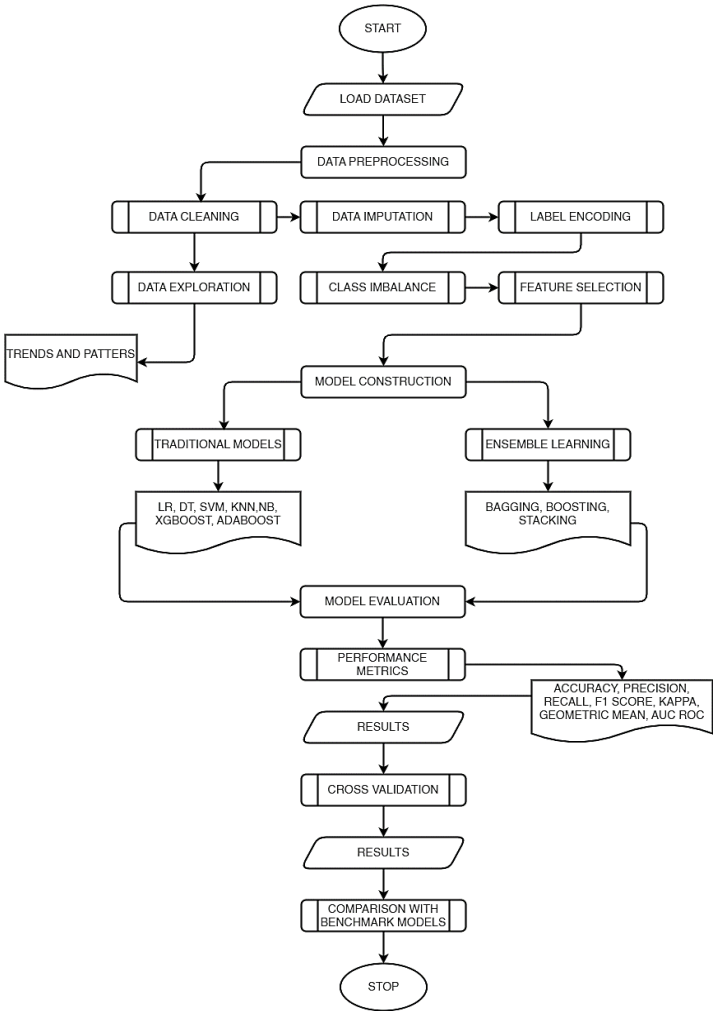

Figure 1. Workflow diagram illustrating experimental procedure

2.4. Model

In pursuit of the overarching goal of assessing the predictive capabilities of ML techniques, the primary objective was to evaluate how effectively machine learning techniques can predict mental health crises. To achieve this, models were built using an imputed dataset, which ensured the integrity and completeness of the data. This robust foundation allowed for more reliable and accurate predictive modeling.

Consequently, the study thoroughly assesses the potential of machine learning as a valuable tool in mental health crisis prediction.

The predictive models themselves encompassed a selection of widely recognized classifiers, including two boosting classifiers known for their ability to enhance predictive performance [28]. The table below summarizes the sequential procedures taken to generate each predictive model and offers a thorough overview of the approaches used for model construction. The results of these model builds serve as essential instruments for evaluating how well machine learning techniques forecast mental health crises. In the end, the comparative analysis helps determine the most promising path for accurate mental health crisis predictions by illuminating the many approaches with differing degrees of effectiveness.

Algorithm 1 outlines the step-by-step algorithm used for developing the machine learning models in this study. It details the data preprocessing stages such as loading the dataset, handling missing target values, and feature selection. Algorithm 1 also covers the classification workflow, including hyperparameter tuning, model training, prediction, performance evaluation, and resource usage monitoring across different classifiers. This structured approach ensures a comprehensive and reproducible model development process.

Table 1 summarizes the different combinations of feature selection methods, number of selected features, and class imbalance strategies used in our models. The features selected in these models include key predictors such as historical symptom severity (total number and duration of crisis episodes), hospital interactions (unplanned contacts, missed appointments, recent crises), patient age, individual risk indices, and total time since the first hospital visit. As shown in Table 1, both EFSA and RFECV were applied, using either eight or five of these significant features, in combination with class imbalance techniques such as SMOTE and Tomek links.

Algorithm 1. Machine learning model development
Start Algorithm:
1. Load the dataset.
2. Remove instances with a null target variable.
3. Apply feature selection to retain relevant features.
4. Split the data into training and testing sets based on the specified ratio.
5. Initialize an empty dictionary to store classifier results.
6. For each classifier:
   – 6.1. Define the hyperparameter grid for hyperparameter tuning.
   – 6.2. Find the best parameters.
   – 6.3. Store the best parameters for the classifier.
7. For each classifier:
   – 7.1. Initialise the classifier with the best hyperparameters.
   – 7.2. Start the timer and memory monitor.
   – 7.3. Train the classifier on the training data.
   – 7.4. Make predictions on the test set.
   – 7.5. Record the elapsed time and memory usage.
   – 7.6. Calculate performance metric
   – 7.7. Store the results in the dictionary with the classifier name as the key.
8. Display the results from the dictionary for each classifier, including:
   – Classifier name
   – Best hyperparameters
   – Performance metrics
   – Elapsed time and memory usage
End Algorithm.

2.5. Evaluation

We approached the crisis prediction task as a binary classification problem [29]. The model was designed to predict the risk of a crisis developing within the next 28 days. To evaluate this, we used a time-based split of the data: 80% for training, 10% for validation, and 10% for testing. The models developed in this study was assessed based on how accurately and effectively they identify patients at risk of a mental health crisis within 28 days following their hospitalization. Model efficacy was measured using a variety of performance indicators, including accuracy, precision, recall, F1 score, and AUC-ROC. Accuracy will provide a broad overview of the model's performance, whilst precision and recall will provide information about the model's capacity to accurately distinguish true positives vs false positives. The F1 score will function as a balanced assessment, which is especially relevant in healthcare settings where false negatives can have serious

repercussions. The AUC-ROC will help illustrate the trade-offs between sensitivity and specificity across different threshold settings, allowing for a more nuanced understanding of model performance.

Table 1. Methods and techniques used to build models

Model	Feature selection method	Number of features	Class imbalance
1	EFSA	8	None
2	EFSA	8	SMOTE
3	EFSA	8	Tomek links
4	RFECV	5	None
5	RFECV	5	SMOTE
6	RFECV	5	Tomek links

3. RESULTS AND DISCUSSION

3.1. Features contributing to mental health crises

The shapley additive explanations (SHAP) summary plot in Figure 2 illustrates the relative importance and directional impact of each feature on the model's prediction of mental health crisis risk. Features are ranked by their overall influence, with "Days since last drug failure," "Age," and "Weeks since last crisis" emerging as the most significant predictors. High values for these features (indicated in red) are associated with an increased predicted risk of crisis, while low values (blue) tend to decrease risk. Notably, variables such as "Not diagnosed," "Days since risk of suicide identified," and "Weeks since last crisis episode" also contribute meaningfully to the model's output. In contrast, features like "Never hospitalized," "Never needed MHA," and "Days since risk of substance misuse identified" have minimal impact, as reflected by their short SHAP bars.

The present findings indicate that the most predictive features for mental health crisis risk closely align with the observations made by [29], Specifically, factors such as the historical severity of symptomsincluding the total number of crisis episodes and the duration of the last episodealong with interactions with the hospital, like unplanned contacts, missed appointments, or a recent crisis, were crucial predictors. Additionally, patient characteristics such as age, individual risk indices, and the total time since the patient's first hospital visit significantly contributed to the model's predictive power.

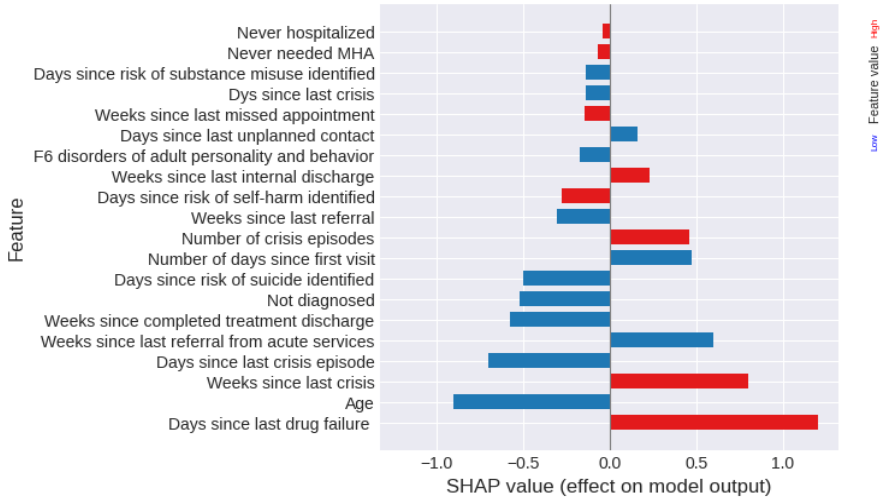

Figure 2. The shapley additive explanations summary plot

3.2. Model 1 with EFSA features and hyperparameter tuning

Model 1, shown in Table 2 constructed without class imbalance techniques, used EFSA-selected features, multivariate imputation by chained equations (MICE)-imputed data, and hyperparameter tuning. The model achieved higher accuracy, precision, recall, and F1 scores across classifiers, reflecting improved discriminatory power as illustrated in Table 2. EFSA identified critical features, while MICE ensured robust data quality. These strategies feature selection, imputation, and tuning collectively outperformed baselines, underscoring their efficacy in enhancing predictions despite omitting class imbalance adjustments. The combination of EFSA and hyperparameter tuning shows notable performance with SVM and LR classifiers achieving 0.88 accuracy.

Table 2. Model 1 performance with EFSA features and hyperparameter tuning

Performance metrics	Classifiers				Boosting classifiers	
	SVM	DT	LR	K-NN	XGBoost	AdaBoost
Accuracy	0.88	0.91	0.88	0.88	0.92	0.92
Precision	0.89	0.93	0.86	0.86	0.91	0.91
Recall	0.96	0.94	0.96	0.98	0.97	0.97
F1 Score	0.92	0.93	0.92	0.91	0.94	0.94
Roc Auc	0.86	0.90	0.86	0.83	0.89	0.89
Kappa	0.75	0.80	0.75	0.71	0.81	0.81
Geometric mean	0.85	0.90	0.85	0.82	0.89	0.89
Balanced Acc	0.86	0.90	0.86	0.83	0.89	0.89
Time (Sec)	0.05	0.006	0.05	0.04	0.49	0.28
CPU (KB)	0	0	0	0	0	0

3.3. Model 2 with SMOTE, EFSA features, and hyperparameter tuning

Model 2 used SMOTE to balance classes, alongside EFSA feature selection and hyperparameter tuning. Results, as shown in Table 3, show classifiers achieved strong accuracy, precision, and recall with balanced data. Hyperparameter tuning again proved effective. SMOTE's synthetic instances risked misrepresenting data distributions, potentially limiting generalizability, but enabled robust minority-class handling. The Model 1 vs. 2 comparison highlights trade-offs: class balancing improved fairness, while tailored attributes optimized performance. Despite minor metric dips, Model 2's integrated approach SMOTE, EFSA, and tuning underscores the value of balancing techniques in mental health prediction, even with inherent compromises, and this validates [9], that SMOTE does not increase accuracy in health predictions since sizes does not count.

Table 3. Model 2 performance with SMOTE, EFSA features, and hyperparameter tuning

Performance metrics	Classifiers				Boosting classifiers	
	SVM	DT	LR	K-NN	XGBoost	AdaBoost
Accuracy	0.89	0.90	0.89	0.87	0.92	0.91
Precision	0.89	0.92	0.86	0.88	0.93	0.92
Recall	0.96	0.92	0.96	0.92	0.96	0.93
F1 Score	0.92	0.92	0.92	0.90	0.94	0.93
Roc Auc	0.86	0.89	0.86	0.84	0.91	0.89
Kappa	0.75	0.77	0.75	0.69	0.82	0.79
Geometric mean	0.85	0.88	0.85	0.84	0.91	0.89
Balanced Acc	0.86	0.88	0.86	0.84	0.91	0.89
Time (Sec)	0.07	0.006	0.05	0.02	0.21	0.56
CPU (KB)	0	0	0	0	264	0

3.4. Model with tomek links, EFSA features, and hyperparameter tuning

Model 3 uses a distinct class imbalance technique with Tomek links. Boosting classifiers, especially XGBoost, perform well, achieving 93% accuracy and 91% balanced accuracy as presented in Table 4. Compared to previous models, Model 3 shows slightly higher accuracy than Model 2. Though Tomek links don't fully balance classes, they highlight the importance of tailored imbalance handling, despite potential information loss in the "No" class affecting generalization.

Table 4. Model 3 performance with Tomek links, EFSA features, and hyperparameter tuning

Performance metrics	Classifiers				Boosting classifiers	
	SVM	DT	LR	K-NN	XGBoost	AdaBoost
Accuracy	0.89	0.92	0.89	0.87	0.93	0.92
Precision	0.89	0.93	0.86	0.87	0.93	0.91
Recall	0.96	0.95	0.96	0.96	0.96	0.98
F1 Score	0.92	0.94	0.92	0.91	0.94	0.94
Roc Auc	0.86	0.90	0.86	0.84	0.91	0.89
Kappa	0.75	0.81	0.75	0.71	0.83	0.82
Geometric mean	0.85	0.90	0.85	0.82	0.91	0.89
Balanced Acc	0.86	0.90	0.86	0.84	0.91	0.89
Time (Sec)	0.09	0.01	0.09	0.04	0.27	0.28
CPU (KB)	0	0	0	0	0	0

3.5. Model with RFECV features and hyperparameter tuning

Model 4, like Model 1, does not apply any class imbalance technique and uses hyperparameter tuning, but distinguishes itself by employing RFECV for feature selection. This approach investigates the

effect of optimal feature selection on model performance, as shown in Table 5. Model 4 achieves high accuracy, precision, and recall (accuracies from 0.89 to 0.93) and this is consistent with [27] that RFECV yields greater accuracy in mental health predictions. While RFECV may reduce feature space and risk information loss, its refined selection clearly boosts overall model effectiveness for mental health prediction.

Table 5. Model 4 performance with RFECV features and hyperparameter tuning

Performance metrics	Classifiers				Boosting classifiers	
	SVM	DT	LR	K-NN	XGBoost	AdaBoost
Accuracy	0.93	0.93	0.89	0.92	0.93	0.92
Precision	0.92	0.92	0.86	0.92	0.92	0.91
Recall	0.97	0.97	0.96	0.96	0.97	0.98
F1 Score	0.95	0.95	0.92	0.94	0.95	0.94
Roc Auc	0.91	0.91	0.86	0.90	0.91	0.90
Kappa	0.83	0.83	0.75	0.82	0.83	0.82
Geometric mean	0.91	0.91	0.85	0.90	0.91	0.89
Balanced Acc	0.91	0.91	0.86	0.90	0.91	0.90
Time (Sec)	0.03	0.004	0.007	0.02	0.06	0.25
PU (KB)	0	0	0	0	0	0

3.6. Model with SMOTE, RFECV features, and hyperparameter tuning

Model 5 mirrors Model 2 by using SMOTE and hyperparameter tuning but further narrows its focus to just three key RFECV-selected features. This streamlined approach highlights the impact of SMOTE resampling and a compact feature set. As shown in Table 6, most classifiers improved, with SVM, DT, K-NN, and XGBoost achieving up to 0.93 accuracy. While this method boosts performance over baselines, the limited features and synthetic data may limit adaptability and generalizability.

Table 6. Model 5 performance with SMOTE, RFECV features and hyperparameter tuning

Performance metrics	Classifiers				Boosting classifiers	
	SVM	DT	LR	K-NN	XGBoost	AdaBoost
Accuracy	0.93	0.93	0.89	0.92	0.93	0.92
Precision	0.92	0.92	0.86	0.92	0.92	0.91
Recall	0.97	0.97	0.96	0.96	0.97	0.98
F1 Score	0.95	0.95	0.92	0.94	0.95	0.94
Roc Auc	0.91	0.91	0.86	0.90	0.91	0.90
Kappa	0.83	0.83	0.75	0.82	0.83	0.82
Geometric mean	0.91	0.91	0.85	0.90	0.91	0.89
Balanced Acc	0.91	0.91	0.86	0.90	0.91	0.90
Time (Sec)	0.03	0.004	0.007	0.02	0.06	0.25
CPU (KB)	0	0	0	0	0	0

3.7. Model with tomek links, RFECV features and hyperparameter tuning

Model 6 in Table 7 adopts the same strategy as Model 3, using Tomek links for class imbalance, hyperparameter tuning, and RFECV for feature selection. Its key distinction lies in the specific features chosen. The combination of Tomek links and RFECV-selected features leads to notable performance gains, with SVM and DT classifiers achieving 0.93 accuracy. Model 6 shows clear improvements across accuracy, precision, and recall, highlighting the synergy of class imbalance handling and targeted feature selection. Table 8 shows the results of the algorithms with Tomek links, RFECV features, and hyperparameter tuning.

Table 7. Model 6 performance with Tomek links, RFECV features, and hyperparameter tuning

Performance metrics	Classifiers				Boosting classifiers	
	SVM	DT	LR	K-NN	XGBoost	AdaBoost
Accuracy	0.93	0.93	0.89	0.92	0.93	0.92
Precision	0.93	0.92	0.86	0.92	0.92	0.91
Recall	0.97	0.97	0.96	0.96	0.97	0.98
F1 Score	0.95	0.95	0.92	0.94	0.95	0.94
Roc Auc	0.91	0.91	0.86	0.90	0.91	0.90
Kappa	0.83	0.3	0.75	0.82	0.83	0.82
Geometric mean	0.91	0.91	0.85	0.90	0.91	0.89
Balanced Acc	0.91	0.91	0.86	0.90	0.91	0.90
Time (Sec)	0.04	0.003	0.007	0.02	0.05	0.25
CPU (KB)	0	0	0	0	0	0

A comprehensive evaluation of various model configurations highlights the importance of tailored machine learning techniques for mental health prediction. Approaches such as SMOTE, Tomek links, EFSA, and RFECV all contributed to improved performance, with hyperparameter tuning further boosting results. SVM and DT classifiers consistently achieved high accuracy, peaking at 93%. However, potential drawbacks include information loss from feature selection and limitations of imbalance techniques. Cross-validation proved essential for assessing model robustness and generalizability. XGBoost stood out as the top performer across most metrics, while AdaBoost excelled in Model 4. SVM and DT also delivered strong results, nearly matching XGBoost's proficiency.

3.8. Ensemble learning

The results obtained when applying the three commonly used EL methods—bagging, boosting, and stacking are presented in Table 8. Ensemble methods are machine learning techniques that combine the predictions from multiple models to improve overall accuracy and performance. Bagging focuses on reducing variance by aggregating results from several models trained on different data samples, while boosting aims to reduce bias by sequentially correcting errors from previous models. Stacking further enhances performance by training a final model to best combine the outputs of various individual models, providing a comprehensive approach to predictive modelling.

3.8.1. Bagging

The bagging EL method was implemented using three diverse classifiers: DT, SVM, and K-NN. Each classifier individually performed well, and their algorithmic diversity enhanced the ensemble's overall robustness and accuracy. As shown in Table 8, the bagging model achieved an accuracy of 0.927, precision of 0.936, and recall of 0.955. The F1 score reached 0.945, and the ROC AUC was 0.913, indicating strong discriminatory power. The kappa coefficient (0.835), geometric mean (0.912), and balanced accuracy (0.913) further confirm model stability. With a training time of just 0.151 seconds and low memory usage, this bagging-SMOTE approach proves both effective and computationally efficient for mental disorder prediction.

3.8.2. Boosting

The boosting EL method was implemented using DT and AdaBoost, resulting in outstanding performance that surpassed other classifier combinations. DT served as a strong base model, while AdaBoost iteratively corrected its errors, leading to enhanced results AdaBoost achieved 92% accuracy and DT 93%. This approach leveraged both model diversity and boosting synergy for greater resilience. The method was tested with both SMOTE and Tomek links for class imbalance, each yielding distinct results while the ROC curve remained consistently high (0.91). Overall, the boosting ensemble method demonstrated superior precision, recall, and robust discriminatory power for mental disorder prediction.

3.8.3. Stacking

The stacking EL method integrates a range of high-performing base classifiers-DT, SVM, K-NN, and XGBoost-each contributing unique algorithmic perspectives and strengths. By combining their predictions through a meta-learner, the stacking ensemble effectively leverages this diversity to enhance overall predictive performance. Results, detailed in Table 8, highlight the method's ability to capitalize on the strengths of each classifier. Achieving a robust ROC AUC of 91%, the stacking approach demonstrates strong, reliable model performance, comparable to the Boosting EL method.

When comparing the EL methods presented above, distinct patterns in their performance are observed with each EL method demonstrating its unique advantages. As observed in Table 8, the performance metrics of different EL methods (SMOTE Bagging, SMOTE Boosting, SMOTE Stacking, and Tomek links Boosting) are displayed. Each method employs various classifiers and techniques to combine their predictions, aiming to enhance overall model performance.

Based on the presented metrics, SMOTE Bagging, SMOTE Boosting, and SMOTE Stacking all yield the same accuracy, precision, recall, F1 score, ROC AUC, geometric mean, and balanced accuracy values of 0.91. Tomek links Boosting lags slightly behind with an accuracy of 0.93, precision of 0.92, recall of 0.97, F1 score of 0.95, ROC AUC of 0.91, and lower kappa of 0.83, as shown in Table 8. Considering the results of models 1 to 5, where it was observed that boosting classifiers tended to perform better across various models and datasets, it can be inferred that SMOTE Boosting might be the most effective EL method among the four. This is because boosting methods excel in correcting errors in base models and using their collective strengths. The SMOTE Boosting method achieved comparable metrics with other methods, such as Bagging and Stacking, requiring minimal training time (0.01 seconds) and negligible memory usage. Our findings resonate with [9], who concluded that ensemble-learning methods boost accuracy in mental health predictions.

While SMOTE Boosting demonstrates strong overall performance across various metrics, there are situations where Tomek links Boosting might be preferred. In the real world, the dataset is not always balanced. If the dataset is particularly sensitive to class imbalance and focuses on accurately predicting the minority class (e.g., mental health), Tomek links Boosting might be more suitable. It showed the highest recall (0.97) among the compared EL methods, indicating its effectiveness in correctly identifying instances of the minority class.

Table 8. Ensemble learning results

Performance metrics	Ensemble learning			
	SMOTE Bagging	SMOTE Boosting	SMOTE Stacking	Tomek Links Boosting
Accuracy	0.93	0.93	0.93	0.93
Precision	0.94	0.94	0.94	0.92
Recall	0.96	0.96	0.96	0.97
F1 Score	0.95	0.95	0.95	0.95
Roc AUC	0.91	0.91	0.91	0.91
Kappa	0.84	0.84	0.84	0.83
Geometric mean	0.91	0.91	0.91	0.91
Balanced Acc	0.91	0.91	0.91	0.91
Time (Sec)	0.15	0.01	0.78	0.02
CPU (KB)	0	0	0	0

3.9. Confusion matrix for the best model

When it is essential to avoid missing a true positive-for instance, in situations like healthcare or safety-sensitive environments-the high recall score achieved by Tomek Links Boosting is particularly advantageous, as shown in Table 8. This approach effectively identifies the majority of true cases within the minority class. Figure 3 presents the corresponding confusion matrix, detailing the model's predictions: 64,891 true positives (correctly identified mental health crises, 29,678 true negatives, 987 false positives, and only 111 false negatives. Minimizing false negatives is essential, as missing individuals needing assistance can have serious consequences. While Tomek Links Boosting may have slightly lower accuracy and precision than other methods, its balanced accuracy and geometric mean are comparable, indicating a strong trade-off between class representation and overall performance. Removing near-neighbor instances via Tomek links sharpens decision boundaries, potentially improving generalization and reducing overfitting, especially with complex or noisy data. Overall, the ensemble methods evaluated demonstrate robust predictive power and practical efficiency for mental disorder prediction. While SMOTE Boosting remains a strong option, this study favours Tomek Links Boosting due to its exceptional 97% recall, aligning with the critical goal of accurately identifying those needing mental health intervention and support.

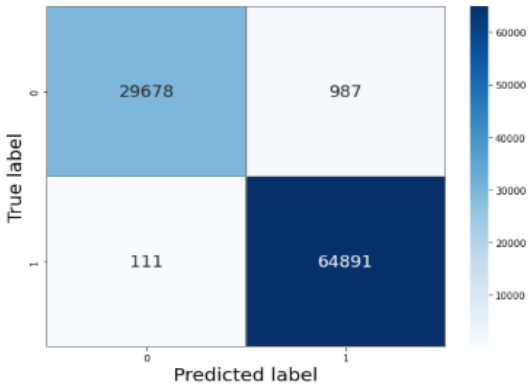

Figure 3. Confusion matrix for Tomek links boosting ensemble learning

4. CONCLUSION

This study provides a comprehensive comparison of ML techniques for predicting mental health crises using a large, longitudinal dataset. Ensemble methods, particularly XGBoost and TlBE, demonstrated superior performance in accuracy, recall, and balanced accuracy. Key risk factors identified include prior hospitalizations, medication adherence, and recent behavioral indicators. These findings suggest that integrating machine learning models into mental healthcare could significantly enhance early identification and intervention for individuals at risk.

ACKNOWLEDGMENTS

We extend our sincere appreciation to the Ministry of Health and Child Care of Zimbabwe for granting us access to the dataset that was instrumental to the success of this project. We appreciate their commitment to advancing research in mental health and their willingness to facilitate academic inquiry through data sharing. This partnership has significantly contributed to the quality and relevance of our findings.

FUNDING INFORMATION

The authors declare they received no funding for this work.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hassan Chigagure		✓		✓		✓	✓	✓	✓		✓			✓
Lucy Charity Sakala	✓	✓			✓					✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This research adheres to all applicable national regulations and institutional policies, aligning with the Ministry of Health and Child Care's declaration. The Harare Institute of Technology institutional review board also approved it.

DATA AVAILABILITY

The data underpinning this study was obtained from Zimbabwe's Ministry of Health and Child Care. Please note that restrictions apply to the availability of this data, as it was used under license for the purpose of this research. Further information can be found through www.mohcc.gov.zw

REFERENCES

[1] World Health Organization, *The World Mental Health Report: transforming mental health for all*. Geneva, Switzerland: WHO, 2022.
[2] R. Filip, R. Gheorghita Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global challenges to public health care systems during the COVID-19 pandemic: A review of pandemic measures and problems," *Journal of Personalized Medicine*, vol. 12, no. 8, 2022, doi: 10.3390/jpm12081295.
[3] J. Lin and W. Guo, "The research on risk factors for adolescents' mental health," *Behavioral Sciences*, vol. 14, no. 4, 2024, doi: 10.3390/bs14040263.
[4] R. Bhatia and J. Kour, "Enhancing social integration and mental health of children with intellectual and developmental disabilities (IDD) through educational programs," *Social Inclusion Tactics for People with Intellectual and Developmental Disabilities*, pp. 321–342, 2024, doi: 10.4018/979-8-3693-3176-7.ch014.
[5] A. Żółtaszek, "WHO bears the high costs of mental health problems in the Labour Force?," *Lodz Economics Working Papers*, 2024.
[6] Indrawati, S.M., Anas, T., Mulyani, S. and Indrawati, N., 2022. *Keeping Indonesia Safe from the COVID-19 Pandemic*. Jakarta, Indonesia: ISEAS-Yusof Ishak Institute.
[7] A. Rembisz and T. Shahal, "Mental health crisis," in *Managing Mental Illness after COVID-19 Infection*, S. A. Collier, New Jersey: John Wiley & Sons, Inc., 2024, pp. 169–193, doi: 10.1002/9781394250103.ch9.
[8] S. C. Faber, A. Khanna Roy, T. I. Michaels, and M. T. Williams, "The weaponization of medicine: Early psychosis in the black community and the need for racially informed mental healthcare," *Frontiers in Psychiatry*, vol. 14, 2023, doi: 10.3389/fpsyt.2023.1098292.
[9] R. Garriga *et al*., "Machine learning model to predict mental health crises from electronic health records," *Nature Medicine*, vol. 28, no. 6, pp. 1240–1248, 2022, doi: 10.1038/s41591-022-01811-5.

[10] D. J. Stein *et al*., "Psychiatric diagnosis and treatment in the 21st century: paradigm shifts versus incremental integration," *World Psychiatry*, vol. 21, no. 3, pp. 393–414, 2022, doi: 10.1002/wps.20998.
[11] L. A. R. Williams and G. Oswald, "Fundamentals of case and caseload management: Skills for rehabilitation practice," *Fundamentals of Case and Caseload Management: Skills for Rehabilitation Practice*, pp. 1–363, 2024, doi: 10.1891/9780826159632.
[12] R. Kotha, P. S. Boyapati, J. Hallur, and V. S. S. R. N. Reddy, "Technological interventions in security and healthcare," *International Journal of Computing and Engineering*, vol. 7, no. 1, 2025.
[13] S. Aminizadeh *et al*., "Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service," *Artificial Intelligence in Medicine*, vol. 149, 2024, doi: 10.1016/j.artmed.2024.102779.
[14] D. Van Sleeuwen *et al.*, "Prediction of long-term physical, mental, and cognitive problems following critical illness: development and external validation of the PROSPECT prediction model," *Critical Care Medicine*, vol. 52, no. 2, pp. 200–209, 2024, doi: 10.1097/CCM.0000000000006073.
[15] J. Chung and J. Teo, "Mental health prediction using machine learning: taxonomy, applications, and challenges," *Applied Computational Intelligence and Soft Computing*, 2022, 2022, doi: 10.1155/2022/9970363.
[16] C. R. Butler, L. B. Webster, and D. S. Diekema, "Staffing crisis capacity: a different approach to healthcare resource allocation for a different type of scarce resource," *Journal of Medical Ethics*, vol. 50, no. 9, pp. 647–649, 2024, doi: 10.1136/jme-2022-108262.
[17] G. Gutierrez, C. Stephenson, J. Eadie, K. Asadpour, and N. Alavi, "Examining the role of AI technology in online mental healthcare: opportunities, challenges, and implications, a mixed-methods review," *Frontiers in Psychiatry*, vol. 15, 2024, doi: 10.3389/fpsyt.2024.1356773.
[18] A. M. Alhuwaydi, "Exploring the role of artificial intelligence in mental healthcare: current trends and future directions – a narrative review for a comprehensive insight," *Risk Management and Healthcare Policy*, vol. 17, pp. 1339–1348, 2024, doi: 10.2147/RMHP.S461562.
[19] J. Zhang, "Understanding the usage of health services by patients with alcohol use disorders," *Doctor of Philosophy Thesis*, School of Computing and Information Technology, University of Wollongong, Wollongong, Australia, 2024.
[20] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowledge and Information Systems*, vol. 66, no. 3, pp. 1575–1637, 2024, doi: 10.1007/s10115-023-02010-5.
[21] M. Nssibi, G. Manita, and O. Korbaa, "Advances in nature-inspired metaheuristic optimization for feature selection problem: a comprehensive survey," *Computer Science Review*, vol. 49, 2023, doi: 10.1016/j.cosrev.2023.100559.
[22] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, 2022, doi: 10.3389/fbinf.2022.927312.
[23] X. Cheng, "A comprehensive study of feature selection techniques in machine learning models," *Insights in Computer, Signals and Systems*, vol. 1, no. 1, pp. 65–78, 2024, doi: 10.70088/xpf2b276.
[24] K. Maharana, S. Mondal, and B. Nemade, "A review: data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltp.2022.04.020.
[25] A. Rehman, S. Naz, and I. Razzak, "Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities," *Multimedia Systems*, vol. 28, no. 4, pp. 1339–1371, 2022, doi: 10.1007/s00530-020-00736-8.
[26] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 598–608, 2023, doi: 10.11591/ijeecs.v29.i1.pp598-608.
[27] T. Mokheleli, "A comparison of machine learning techniques for predicting mental health disorders," *Master of Commerce Thesis*, in Information Technology Management, Department of Applied Information Systems, College of Business and Economics, University of Johannesburg, Johannesburg, South Africa, 2023.
[28] M. U. Bokhari, G. Yadav, Zeyauddin, and S. Afzal, "Enhancing mental health prognosis: an investigation of advanced hybrid classifiers with cutting-edge feature engineering and fusion strategies," *International Journal of Information Technology*, 2024, doi: 10.1007/s41870-024-02092-6.
[29] J. Chung and J. Teo, "Single classifier vs. ensemble machine learning approaches for mental health prediction," *Brain Informatics*, vol. 10, no. 1, 2023, doi: 10.1186/s40708-022-00180-6.

BIOGRAPHIES OF AUTHORS

Hassan Chigagure holds a Bachelor of Science degree in Operations Research from the National University of Science and Technology, Zimbabwe (2023). He is presently working towards an M.Tech. in Data Science and Analytics with Harare Institute of Technology and an M.Sc. in Financial Engineering with National University of Science and Technology Hassan is passionate about financial modeling, big data applications in finance, and quantitative risk modeling. He can be contacted at email: hassanchigagure@icloud.com.

Lucy Charity Sakala received her first degree from Bindura University of Science Education (BUSE), Computer Science in 2007. She has a Master's degree in Computer Science, University of Zimbabwe (UZ, 2011). The Ph.D. in Information Systems degree from the University of Cape Town (UCT) South Africa, 2019. She is currently a lecturer in Computer Science and Information Technology. Her research interests are in artificial intelligence, human computer interaction, information systems, software engineering, ICT for development, and Digital transformation. She is a member of Computer Society of Zimbabwe and Zimbabwe. She has over 10 publications and Scopus indexed book chapter. She can be contacted at email: lsakala@buse.ac.zw, lsakala@hit.ac.zw.