Interpretable Machine Learning Framework for Early Depression Detection Using Socio-Demographic Features with Dual Feature Selection and SMOTE

Zineb Sabouri*[1], Imane Moustati [1], Noreddine Gherabi [1] and Mohamed Amnai [2]
[1]Lasti Laboratory Khouribga, Sultan Moulay Slimane University, ENSA Khouribga, Beni Mellal, Morocco
[2]Laboratory of Computer Sciences, Faculty of Sciences, IbnTofail University, Kenitra, Morocco
E-mail: zineb.sabouri@usms.ac.ma
*Corresponding author

Keywords: depression, SMOTE, ANOVA, ML algorithms, SHAP

Received: July 17, 2025

Depression is the most widespread psychological disorder globally, impacting individuals across all age groups; when left undiagnosed or untreated, it significantly elevates the risk of severe outcomes, including suicidality. This study explores the efficacy of eight machine learning (ML) classifiers utilizing socio-demographic and psychosocial data to discern signs of depression. A depression dataset available on GitHub was acquired, comprising 604 instances with 30 predictors and 1 target variable indicating depression status. Preprocessing included normalization, handling missing values, and encoding categorical variables. Two feature selection methodologies, Analysis of Variance (ANOVA) and Boruta were employed to extract pertinent features. ANOVA selected 19 features, while Boruta retained 13 for model training. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was utilized to enhance prediction accuracy (ACC). Results demonstrate that Logistic Regression (LR), combined with ANOVA feature selection, exhibits superior performance, achieving an ACC of 92.56% and an AUC of 92.69%. With Boruta, LR achieved an ACC of 91.74% and an AUC of 91.65%. Without feature selection, LR yielded an ACC of 87.75%, a precision of 91.73%, and an AUC of 89.98%. SHapley Additive exPlanations (SHAP) analysis revealed that anxiety (ANXI) is the most influential predictor within the ML model designed for depression prediction. This study identifies the most effective model for predicting depression through evaluation metrics, while also addressing societal biases and supporting clinicians with interpretable insights for early intervention.

Povzetek: Raziskan je razložljiv okvir strojnega učenja za zgodnje odkrivanje depresije na podlagi socio-demografskih podatkov. Z uporabo dvojne izbire značilk in uravnoteženja razredov s SMOTE model izboljša točnost napovedi ter hkrati omogoča interpretacijo vpliva posameznih dejavnikov na odločanje modela.

1 Introduction

Depression is recognized as a complex mental health disorder characterized by persistent feelings of sadness, loss of interest in daily activities, and significant impairments in social and occupational functioning. It is a multifaceted condition that not only affects emotional well-being but also interferes with physical health, often contributing to chronic illnesses such as diabetes and cardiovascular disease [1], [2]. The global burden of depression has been exacerbated by the COVID-19 pandemic, with recent estimates indicating that approximately 322 million individuals worldwide are affected [2]. More critically, depression is a leading contributor to suicide, accounting for nearly 50% of all cases annually [2]. Beyond its psychological impact, depression has profound social and economic consequences [3]. Affected individuals frequently withdraw from social interactions, which can impair relationships and potentially lead to unhealthy coping mechanisms such as substance abuse and overeating [4]. Moreover, fatigue and decreased productivity associated with depression can result in reduced workforce efficiency, thereby influencing a nation's socio-economic performance. Early detection of depression is therefore essential, as timely intervention can significantly alleviate psychological distress and associated somatic symptoms, including sleep disorders and gastrointestinal disturbances [5]. In this context, the integration of ML techniques, along with smartphone-based assessment [6] and intelligent cognitive assistants for attitude and behavior change support [7],[8] into mental health diagnostics presents a transformative opportunity [9],[10]. While ML has been extensively applied in various medical fields, such as COVID-19 detection [11], Alzheimer's disease classification [12], and breast cancer diagnosis [13], its application in psychological analysis remains relatively underexplored. This study aims to bridge that gap by proposing a robust ML-based framework for depression prediction. Specifically, we investigate and compare the performance of several ML classifiers; LR, Naive Bayes (NB), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost); on a publicly available dataset. Our approach not only seeks to forecast depression with high ACC but also to identify key psychosocial and socio-demographic features that significantly influence predictive outcomes. We further enrich our analysis with feature-importance visualizations, including correlation heatmaps and ANOVA-based importance charts, and incorporate explainability methods to help clinicians clearly understand and interpret these key predictors.

To guide this study, we formulate the following research question: Does the choice of feature selection method significantly impact the performance and interpretability of ML classifiers in detecting depression using socio-demographic and psychosocial data? We hypothesize that statistical methods like ANOVA improve predictive ACC compared to wrapper-based approaches such as Boruta or models without feature selection.

The primary contributions of this study are threefold:

- Identification of influential psychosocial and socio-demographic variables relevant to depression prediction;
- Construction of optimized, domain-specific datasets for enhanced screening ACC;
- Comprehensive evaluation of multiple ML algorithms and feature selection techniques to develop an efficient and interpretable depression prediction model.
- Integration of SHAP for interpretability and explainability, providing detailed insights into feature importance and the contribution of each variable to model predictions.

The remainder of this paper is structured as follows: Section 2 reviews related literature; Section 3 outlines the proposed methodology; Section 4 presents experimental results and comparative analysis; and the final section discusses conclusions and directions for future research.

2 Related works

This section presents an extensive review of related research, focusing on the methodologies employed, key predictors identified, and algorithmic performance, with the aim of identifying existing research gaps and informing future modeling efforts. In a study involving 84 breast cancer patients, socio-demographic variables and Beck Depression Inventory (BDI) scores were used to evaluate three ML models, with Artificial Neural Networks (ANN), particularly those utilizing extreme learning strategies, achieving the highest ACC [14]. Similarly, in an elderly cohort aged 60 and above, RF outperformed Support Vector Machine (SVM) and LR, achieving an AUROC of 0.83, with brain region volumes, depression symptomatology, and self-reported health-related quality of life emerging as key predictors [15]. Extending this work, our prior research applied six supervised models; kNN, RF, LR, DT, SVM, and NB to a public dataset, revealing that SVM and LR achieved superior performance with an ACC of 83.32% using 10-fold cross-validation [16]. Addressing the issue of class imbalance, a South Korean study employed RF combined with SMOTE to forecast depression onset, achieving 86.20% ACC and identifying socio-familial satisfaction and perceived health as critical features [17]. The promise of deep learning was demonstrated in a study using a multilayer perceptron (MLP) with backpropagation to assess depression in working professionals, yielding a remarkable 98.8% ACC [18]. Complementing these findings, a comprehensive review on bipolar disorder diagnostics highlighted the dominance of classification models especially those using MRI data while noting underutilization of genomic and microarray data [19]. Further evidence supporting ensemble approaches comes from an empirical study that evaluated six ML classifiers with three feature selection methods, identifying AdaBoost with SelectKBest as the most effective combination, reaching an ACC of 92.56% [20]. A more recent investigation applied a suite of models including SVM, kNN, LR, RF, XGBoost, and Neural Networks to jointly predict depression and Generalized Anxiety Disorder (GAD), demonstrating competitive performance and underscoring the feasibility of comorbidity modeling [21]. Moreover, psychometric and demographic predictors have been leveraged using XGBoost, which outperformed traditional LR [22], while studies on postpartum depression (PPD) revealed the superiority of Functional Gradient Boosting in prediction ACC [23]. Notably, the development of a mobile-based Clinical Decision Support System (CDSS) incorporating NB, LR, SVM, and ANN for early PPD detection represents a significant step toward real-time clinical implementation [24]. Collectively, these studies underscore the effectiveness of ensemble and deep learning techniques, particularly RF, XGBoost, and MLP in depression prediction across various demographic and clinical settings. All these studies are summarized in Table 1, which presents for each study the dataset size, feature types, algorithms compared, performance evaluation, and the best-performing algorithm.

Table 1: A comparative summary of State-of-the-Art (SOTA) approaches for depression detection

Study	Dataset Size	Feature Types	Algorithms Compared	Best Performance
[14] J. Cvetković	84 patients	Socio-demographic + BDI scores	– ANN with extreme learning algorithm. – ANN with back propagation learning algorithm. – fuzzy with genetic algorithm.	ANN with extreme learning algorithm.
[15] Grzenda et al.	67 elderly patients (≥60 years) from 2 clinical trials (NCT01902004, NCT02460666)	Structural MRI (GMV), clinical measures, self-reports, cognitive tests, demographics, treatment response	RF, SVMRBF, LR	RF : Test AUC 0.84, MCC 0.47;
[16] Sabouri et al.	1,409 patients	Socio-demographic & psychometric variables (20 attributes)	LR, SVM, KNN, RF, NB, DT	SVM & LR (ACC 83.32%)
[17] K.-S. Na et al.	6,588 patients	Sociodemographic, quality of life, health, altruistic behaviors	RF + SMOTE	RF : AUROC 0.87, ACC 86.2%
[18] Vincent et al.	1,032 patients	Questionnaire-based: sleep patterns, mood, eating interest, weight, happiness, concentration; Sensor-based: heart rate, sleep duration	Deep multilayer perceptron (MLP) with and without backpropagation	Deep-MLP with backpropagation – ACC > 98%
[20] Zulfiker et al.	604 participants	30 socio-demographic + 25 psychosocial factors	KNN, AdaBoost, Gradient Boosting, XGBoost, Bagging, Weighted Voting	AdaBoost + SelectKBest: ACC 92.56%
[21] Nemesure et al.	>2,500 participants	Standard clinical, demographic, and biomedical features	SVM, kNN, LR, RF, XGBoost, Neural Networks	GAD: 66% sensitivity at 70% specificity
[22] Hatton et al.	<200 (training dataset)	Psychometric + demographic	XGBoost vs. LR	XGBoost : AUC 0.72
[23] Natarajan et al.	N = 173 new mothers (25% with PPD symptoms)	Clinical, psychometric	Functional Gradient Boosting, NB, J48, SVM, AdaBoost, Bagging, LR	FGB: ROC 0.952, Precision 0.920, Recall 0.840.
[24] Jiménez-Serrano et al.	1,397 postpartum women from 7 Spanish hospitals	Demographic + psychometric	NB, LR, SVM, ANN	NB (G = 0.73, and ACC ≈ 0.73)

These studies mainly focused on predicting psychiatric outcomes or postpartum depression using limited ML methods, often relying on clinical, imaging, or questionnaire-based data, and rarely considering socio-demographic features. Most studies did not test multiple feature selection methods, apply data balancing techniques such as SMOTE, or evaluate models with a wide range of metrics like precision, recall, and F1-score, leaving a gap in the comprehensive assessment of predictive performance. Our study fills this gap by predicting depression using socio-demographic features and comparing eight ML algorithms (RF, XGBoost, LightGBM, SVM, KNN, DT, LR, NB), with and without feature selection (ANOVA, Boruta), providing a detailed multi-metric evaluation and demonstrating how feature selection and data balancing enhance predictive performance while SHAP was applied to interpret the results and provide clear, user-friendly explanations.

3 Methodology

The methodological framework of this study begins with the acquisition of a publicly available dataset, followed by a comprehensive preprocessing phase involving data cleaning, normalization, and correction of class imbalance. To enhance model interpretability and predictive performance, two feature selection techniques were employed to identify the most relevant variables. Subsequently, multiple supervised ML algorithms were implemented and systematically evaluated using standard performance metrics, including ACC, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (AUC). The complete workflow is structured sequentially and illustrated in Figure 1 to provide a clear overview of the experimental process.

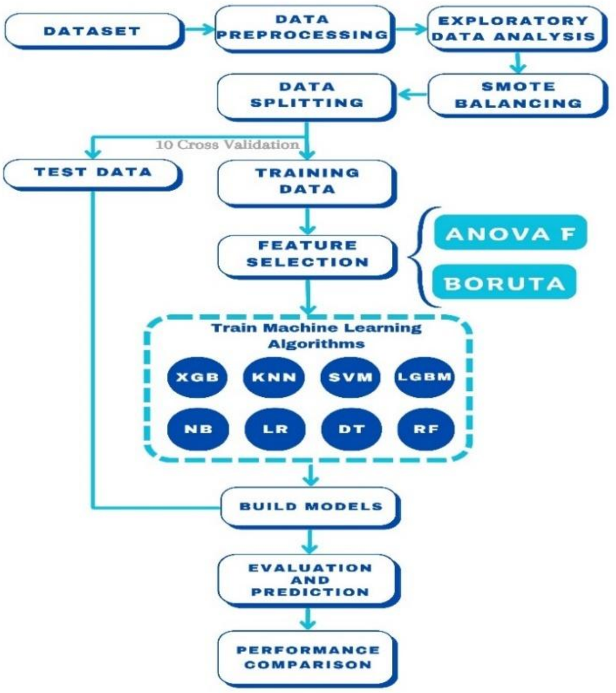

Figure 1: Pipeline of the research methodology

3.1 Dataset description

We employ a publicly available depression screening dataset comprising 604 instances of adults and youths, each annotated with a binary depression label derived from the Depression Checklist. The dataset is available on GitHub [25]. Socio-demographic and psychosocial predictors include age, gender, marital status, employment status, education level, family support score, substance use indicator, and four psychosocial-stress subscales. Table 2 provides a detailed description of each variable in the dataset.

Table 2: Variables description

Variable type	Variable Name	Variable Description
Predictors	AGERNG	Age in years
	GENDER	Gender
	EDU	Educational Attainment
	PROF	Profession
	MARSTS	Marital Status
	RESDPL	Residence Type
	LIVWTH	Whether living with family or not
	ENVSAT	Satisfied or not with the environment
	POSSAT	Whether or not satisfied with current position or achievements
	FINSTR	Financial stress
	DEBT	Had Debt
	PHYEX	Physical Exercise
	SMOKE	Smoker
	DRINK	Alcohol Drinker
	ILLNESS	With Illness
	PREMED	Taking Prescribed Med
	EATDIS	Has Eating disorder
	AVGSLP	Average sleep hours
	INSOM	Has Insomnia
	TSSN	Ave hours in social network
	WRKPRE	Has Work/Study Pressure
	ANXI	Feels anxiety
	DEPRI	Feels deprived
	ABUSED	Felt Abused
	CHEAT	Felt Cheated
	THREAT	Faced threat
	SUICIDE	Suicidal thoughts
	INFER	Inferiority Complex
	CONGLICT	In Conflict with Family or Friends
	LOST	Recent loss of a close person
Target	DEPRESSED	Depressed or not

3.2 Data preprocessing

The original dataset contained 604 instances. After removing 111 records with missing values and 10 duplicate entries, a total of 483 clean instances remained. These were then split into a 70% training set (338 instances) and a 30% test set (145 instances).

3.3 Feature selection

While constructing a ML model, selecting only the essential features is crucial. Including irrelevant features can lead to a decrease in the performance of the model. This phenomenon, known as the "curse of dimensionality," can cause issues such as overfitting, increased computational complexity, and reduced interpretability of the model. Therefore, feature selection techniques are employed to identify and retain only the most informative and relevant features, improving the model's efficiency and ACC. To isolate the most informative predictors and reduce overfitting risk, we apply two complementary filter-based methods independently.

- Analysis of Variance (ANOVA) F-Test: Computes the F-statistic for each continuous or one-hot encoded feature against the binary label. Features with p-values < 0.05 are retained.
- Boruta Algorithm: Employs a randomized wrapper around a RF classifier to iteratively compare original feature importance against shadow (permuted) features. Only attributes with confirmed importance beyond the maximum shadow are selected.

Both approaches are founded on robust statistical principles and computational frameworks, ensuring a thorough and rigorous selection of relevant features for our analysis. Each selection strategy yields a distinct feature subset; downstream models are trained and evaluated separately on these subsets to compare their impact.

3.4 Class imbalance handling

The training data exhibit a depression-to-non-depression ratio of approximately 1:2. To mitigate bias toward the majority class, we apply the Synthetic Minority Oversampling Technique (SMOTE) exclusively on the training fold within each cross-validation iteration, generating synthetic minority samples until class balance is achieved. The test set remains untouched to provide an unbiased evaluation. SMOTE generates synthetic samples of the minority class by interpolating between existing minority instances and their nearest neighbors within the feature space.

3.5 Classifier training and validation

We evaluated eight complementary classification algorithms to predict depression outcomes. The selected models include RF, XGBoost, LightGBM, SVM, kNN, DT, LR, and NB, to ensure a broad survey of both linear and nonlinear decision boundaries.

For each algorithm, we conducted a nested 10-fold cross validation on the training set to optimize hyperparameters: for LR we tuned the regularization coefficient C; for SVM we searched over C and γ; for DT, RF, and XGBoost we varied maximum tree depth, number of estimators, and learning rate (for XGBoost); for kNN we evaluated k values from 3 to 11. Within each outer fold, a grid search over the relevant hyperparameter combinations selected the model configuration that maximized mean area under the ROC curve (AUC). This nested scheme prevents information leakage from the test folds into the tuning process and yields an unbiased estimate of generalization performance. After finalizing hyperparameters, each classifier was retrained on the full training set and then assessed on the independent hold out test set. We report five primary performance metrics: ACC, precision, recall, F1 score, and AUC, to capture both overall correctness and balance-aware behavior. To account for potential class-imbalance effects beyond ACC, we included AUC as a threshold-independent measure of discriminative ability. All experiments were executed with a fixed random seed (42) for data splits, SMOTE sampling, and algorithmic initializations. This diverse selection supports a robust comparative analysis of model performance in depression detection.

3.6 SHAP analysis

The SHAP analysis method is a powerful technique for model interpretability, which borrows the concept of Shapley values from cooperative game theory to fairly explain the output of complex ML models [26]. For a given prediction, SHAP assigns a value to each input feature, representing its unique contribution to the final result compared to the average prediction. This is achieved by calculating the feature's average marginal contribution across all possible groupings of features. This rigorous approach ensures a fair credit allocation among the predictors, thus promoting transparency and building trust by clearly showing how each factor drives specific model decisions [27].

4 Results and discussion

4.1 Results of exploratory data analysis

Prior to applying SMOTE, the dataset exhibited class imbalance, comprising 163 non-depressed instances (33.7%) and 320 depressed instances (66.3%). Following the application of SMOTE, the classes were balanced, with 320 instances in both the non-depressed and depressed categories, facilitating unbiased model training (see Figure 2).

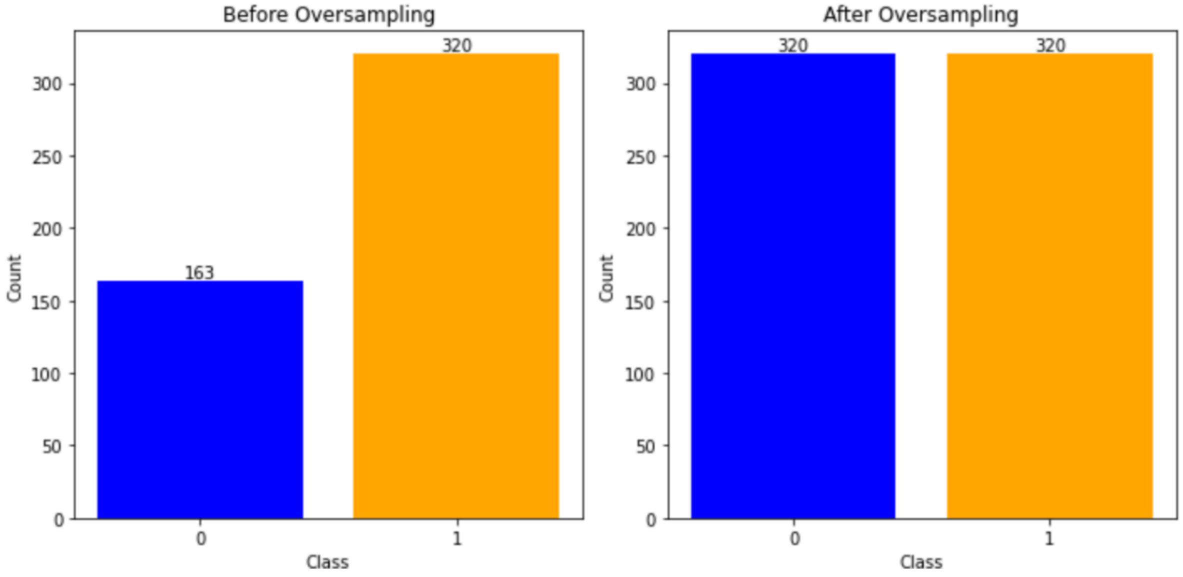

Figure 2: Class distribution before and after SMOTE

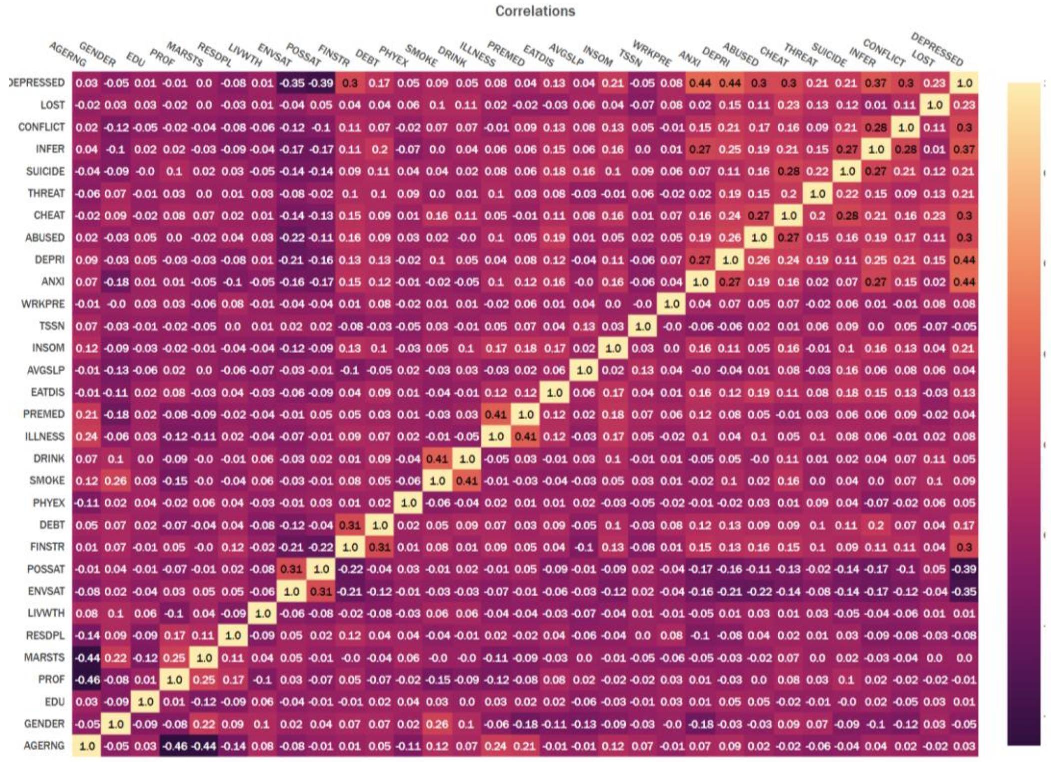

Figure 3: Correlation matrix

A heatmap was generated to visualize correlations among the 31 attributes in the depression dataset (see Figure 3). Notably, several variables including FINSTR, ANX, DEPRI, ABUSED, CHEAT, INFER, and CONFLICT showed significant positive correlations with the depression attribute, with correlation coefficients of 0.3, 0.44, 0.44, 0.3, 0.3, 0.37, and 0.3 respectively. Conversely, variables ENVSAT and POSSAT exhibited negative correlations with depression, with coefficients of -0.35 and -0.39 respectively. On the contrary, attributes such as AGERNG, GENDER, EDU, MARSTS, DRINK, and PREMED demonstrated weak correlations with the target variable.

4.2 Results of the experiment

In this study, we identified and tested 8 distinct ML algorithms: RF, XGBoost, LightGBM, SVM, KNN, DT, LR, and NB. Subsequently, these algorithms were executed both without any feature selection technique and with two feature selection techniques, namely ANOVA and Boruta, to obtain optimal results. This section discusses the outcomes and performance of all the employed classifiers. The table 3 displays the measured ACC, precision, recall, F1-score, and AUC of the classifiers for the constructed models (see Table 3). Notably, the LR classifier yielded the most favorable outcome, achieving an ACC of 87.75%, precision of 91.73%, and an AUC of 89.98% without feature selection. Conversely, KNN exhibited the lowest ACC at 69.21%. Additionally, the accuracies of RF, XGBoost, LightGBM, SVM, DT, and NB classifiers were reported as 86.59%, 84.11%, 84.11%, 85.26%, 74.01%, and 81.79%, respectively.

Notably, applying various feature selection techniques led to a significant improvement in the accuracies of all classifiers. Particularly, when employing the ANOVA feature selection technique, LR emerged as the top-performing classifier across all metrics. It achieved an ACC of 92.56%, a precision of 95.95%, a f1-score of 94.04%, and an AUC of 92.69%. This strong result was further detailed by its ANOVA-derived confusion matrices (Figure 4), which showed 41 true positives, 71 true negatives, 6 false positives, and 3 false negatives. In contrast, the lowest performance was shown by DT with an ACC of 76.03%. The other classifiers namely, RF, XGBoost, LightGBM, SVM, KNN, and NB, are 85.12%,87.60%, 91.74%, 89.26%, 80.17%, and 81.82%, respectively. With the Boruta technique, LR outperforms other classifiers, achieving an ACC of 91.74%. In contrast, the NB classifier achieved the lowest ACC at 80.17%. The accuracies of RF, XGBoost, LightGBM, SVM, KNN, and DT classifiers were reported as 86.78%, 88.43%, 88.43%, 90.91%, 85.12%, and 80.99%, respectively. Upon comparing the outcomes of various models, it is evident that the LR classifier utilizing the ANOVA technique outperformed other models.

The Boruta feature selection algorithm identified 13 predictor variables as irrelevant based on their Maximum Z-score. Meanwhile, the ANOVA technique ranked predictor variables according to their importance, with Table 4 showcasing the top 19 features selected. According to the ANOVA technique (see Figure 5), the most crucial features for depression prediction include ANXI, DEPRI, POSSAT, INFER, and ENVSAT. This difference arises because ANOVA retains a larger set of features, capturing weaker signals, while Boruta focuses only on the strongest predictors. Both approaches highlight variables that align with clinical assessments of depression, underscoring their relevance in predictive modeling.

Table 3: Performance of the classifiers using different feature selection techniques.

ANOVA		**BORUTA**	
Feature	**Importance**	**Feature**	**Score**
ANXI	147.9171	ENVSAT	1
DEPRI	141.488	POSSAT	1
POSSAT	109.1659	FINSTR	1
INFER	94.6283	INSOM	1
ENVSAT	82.1469	ANX	1
CHEAT	59.5334	DEPRI	1
ABUSED	59.383049	ABUSED	1
CONFLICT	58.012737	CHEAT	1
FINSTR	57.723602	THREAT	1
LOST	33.349831	SUICIDE	1
SUICIDE	28.901672	INFER	1
INSOM	27.973808	CONFLICT	1
THREAT	27.499882	LOST	1
DEBT	17.986887		
EATDIS	10.575417		
SMOKE	4.693574		
WRKPRE	4.291038		
RESDPL	3.685119		
ILLNESS	3.415093		

Table 4: The important features

Feature Selection Technique	Classifier Name	ACC	Precision	Recall	F1-score	AUC
No Feature selection	RF	0.8659	0.8970	0.8992	0.8981	0.8506
	XGBoost	0.8411	0.8698	0.8917	0.8806	0.8178
	LightGBM	0.8411	0.8772	0.8816	0.8794	0.8224
	SVM	0.8526	0.9053	0.8665	0.8855	0.8463
	KNN	0.6921	0.9306	0.5743	0.7103	0.7461
	DT	0.7401	0.8191	0.7758	0.7969	0.7237
	LR	**0.8775**	**0.9173**	**0.8942**	**0.9056**	**0.8698**
	NB	0.8179	0.8747	0.8438	0.8590	0.8060
ANOVA	RF	0.8512	0.9041	0.8571	0.8800	0.8490
	XGBoost	0.8760	0.9189	0.8831	0.9007	0.8734
	LightGBM	0.9174	0.9467	0.9221	0.9342	0.9156
	SVM	0.8926	0.9211	0.9091	0.9150	0.8864
	KNN	0.8017	0.9206	0.7532	0.8286	0.8198
	DT	0.7603	0.8529	0.7532	0.8000	0.7630
	LR	**0.9256**	**0.9595**	**0.9221**	**0.9404**	**0.9269**
	NB	0.8182	0.9508	0.7532	0.8406	0.8425
Boruta	RF	0.8678	0.9296	0.8571	0.8919	0.8718
	XGBoost	0.8843	0.9315	0.8831	0.9067	0.8847
	LightGBM	0.8843	0.9315	0.8831	0.9067	0.8847
	SVM	0.9091	0.9342	0.9221	0.9281	0.9042
	KNN	0.8512	0.9538	0.8052	0.8732	0.8685
	DT	0.8099	0.9219	0.7662	0.8369	0.8263
	LR	**0.9174**	**0.9467**	**0.9221**	**0.9342**	**0.9165**
	NB	0.8017	0.9344	0.7403	0.8261	0.8247

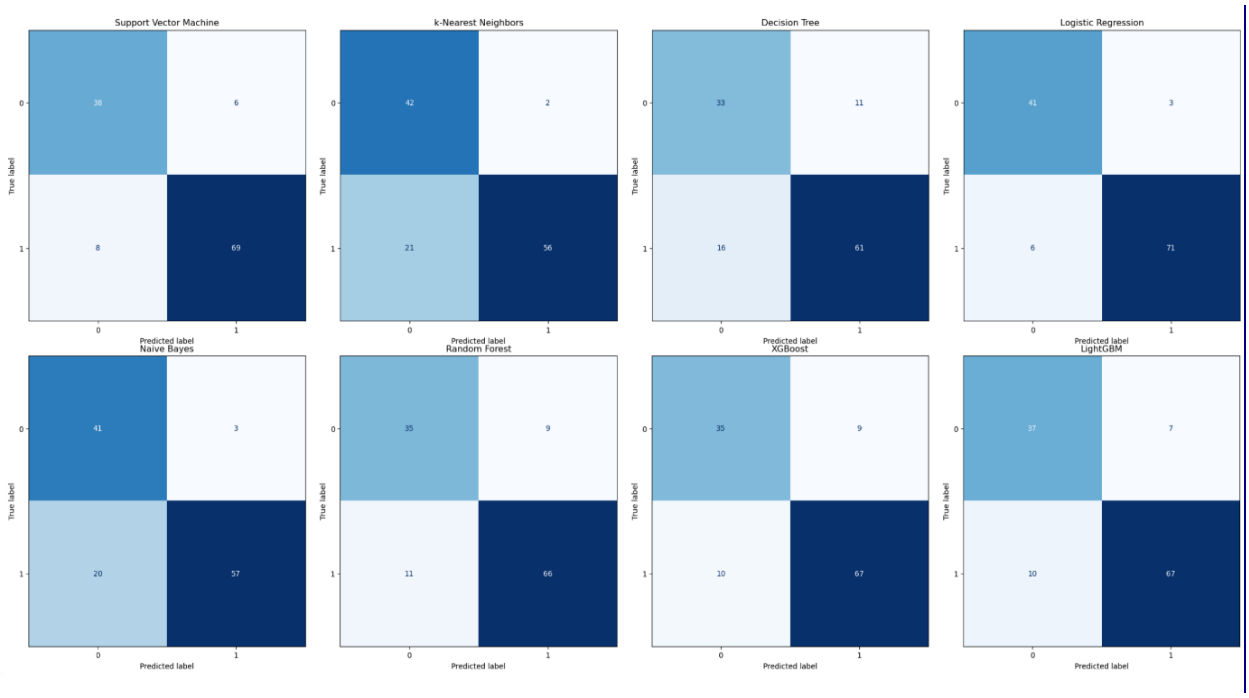

Figure 4: Confusion matrices for all algorithms following ANOVA feature selection

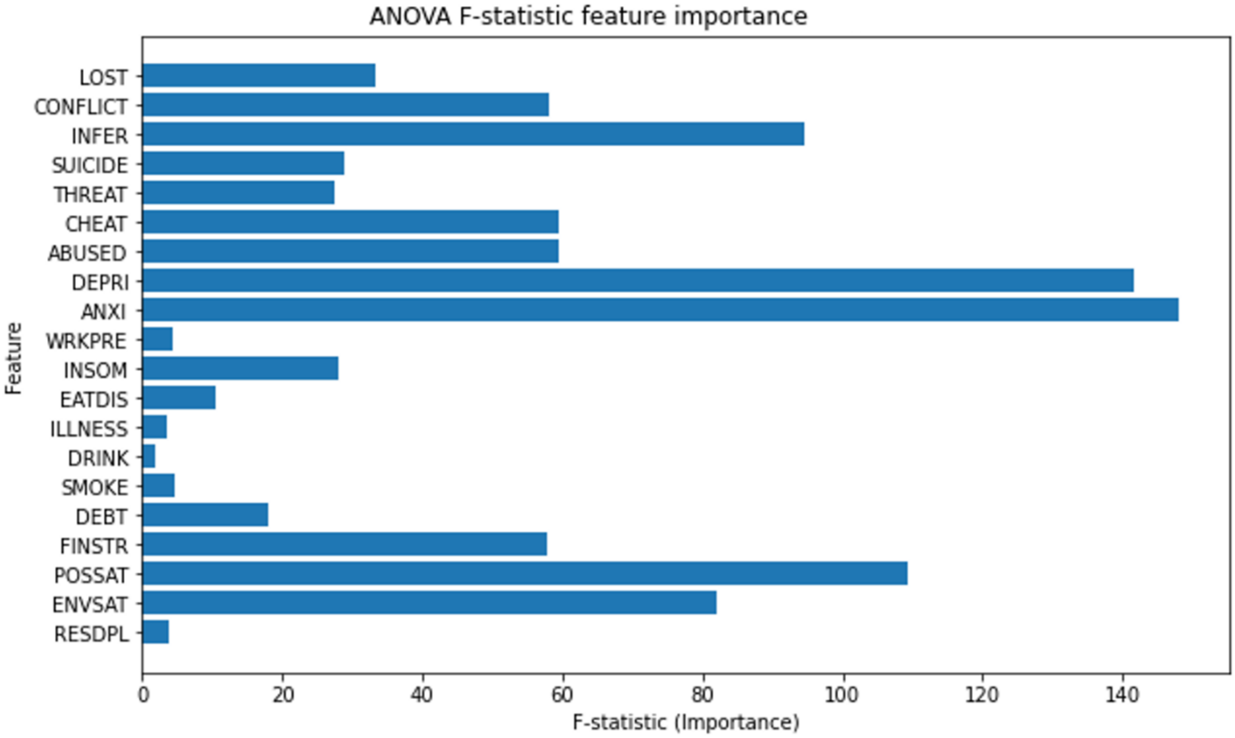

Figure 5: ANOVA F-statistic feature importance

4.3 SHAP interpretability results

Among the eight classification algorithms evaluated, LR achieved the highest performance using the ANOVA-filtered features. To interpret this model, SHAP analysis was applied. The SHAP beeswarm plot (Figure 6) shows that ANXI has the strongest influence on depression prediction, followed in importance by POSSAT, DEPRI, ENVSAT, and INFER. The waterfall plot confirms these findings at the individual case level, indicating that ANXI has a SHAP value of −1.01, exerting a negative influence on the prediction, while LOST contributes positively (+0.89) and INSOM also has a positive effect (+0.72). Conversely, POSSAT (−0.67), DEPRI (−0.65), and ENVSAT (−0.48) contribute negatively (see figure 7).

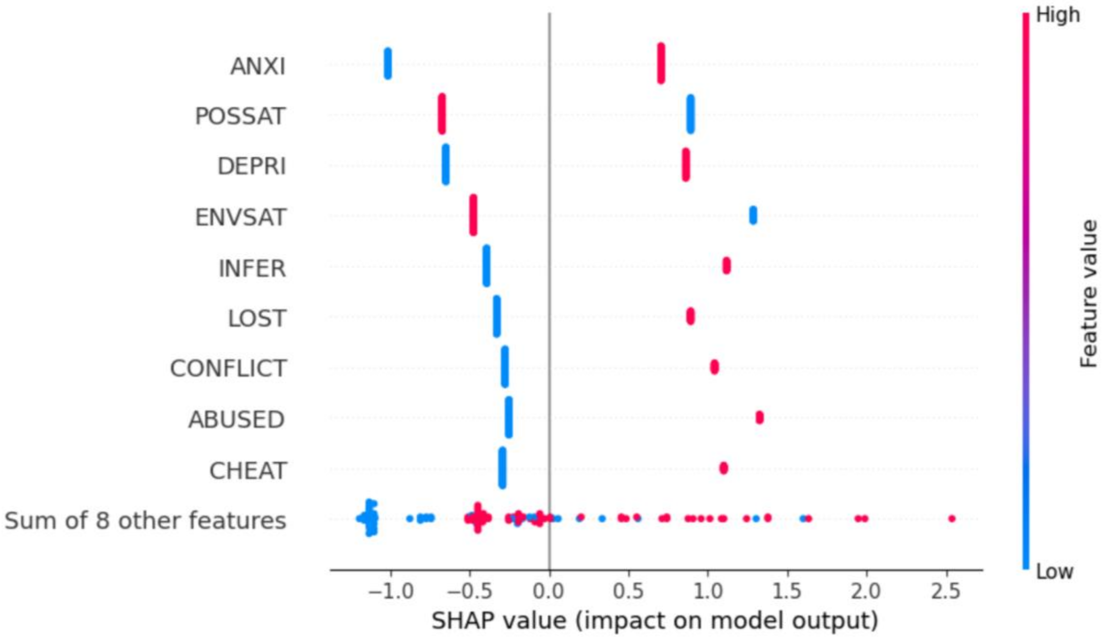

Figure 6: SHAP Beeswarm plot for LR

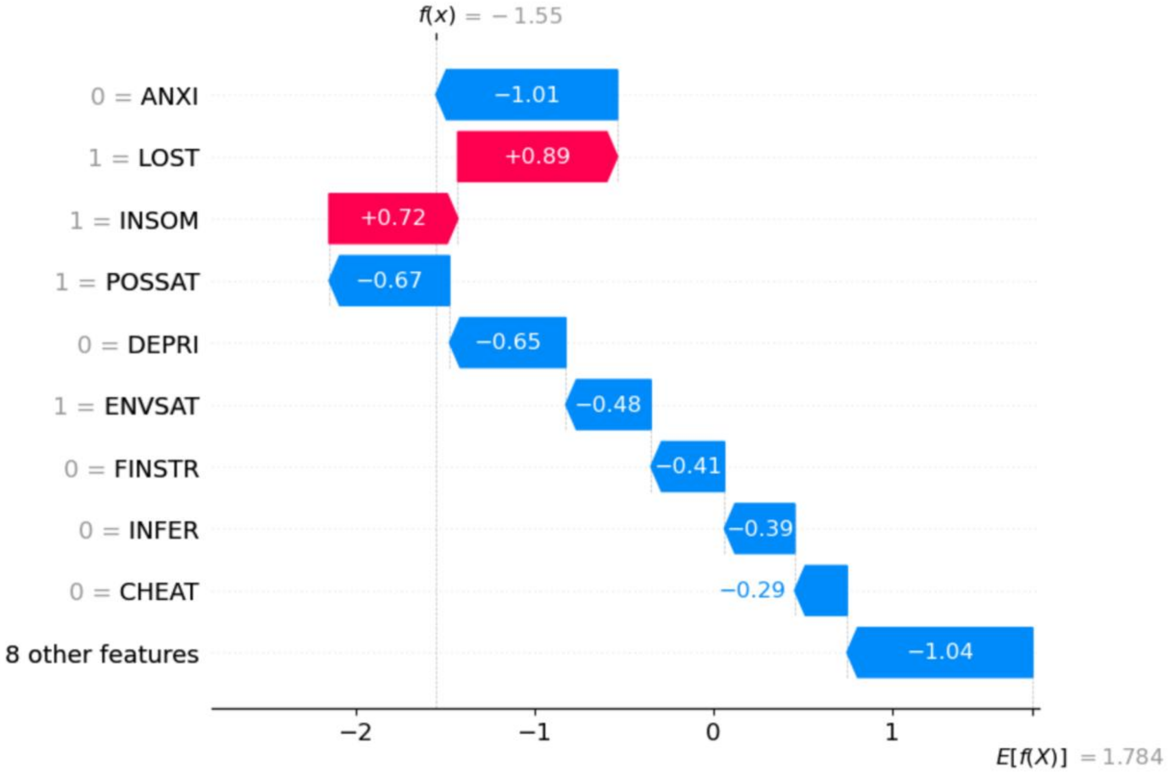

Figure 7: SHAP waterfall plot for LR

5 Discussion

To evaluate the robustness and relevance of this research, it is essential to contextualize its findings within the existing literature. Prior studies in depression prediction have often been constrained by limited demographic scopes, typically focusing on specific subpopulations, such as individuals within a particular age range [15], occupational category [23],[24], or health condition [14]. While these studies have provided valuable insights, they frequently lack generalizability due to their narrow focus. Moreover, many of them primarily aim to detect depression without offering a comprehensive understanding of the contributing factors. In contrast, the present study significantly broadens the scope by incorporating a heterogeneous dataset encompassing individuals from diverse age groups, professions, and socioeconomic backgrounds. This inclusive approach enhances the generalizability and real-world applicability of our findings. Not only did we focus on predicting depression with high ACC, but we also identified and ranked the most influential predictors through rigorous feature selection techniques. Notably, the use of ANOVA for feature selection proved to be both effective and interpretable, reinforcing its utility in mental health research, especially in identifying variables with statistically significant influence on depressive outcomes. Our best-performing ML models achieved consistently high ACC and precision scores, underscoring their potential as effective and scalable screening tools in clinical and public health settings. These models could be particularly valuable in resource-limited environments, enabling early detection and intervention through automated assessments. Furthermore, the integration of SMOTE to address class imbalance represents a methodological strength, improving model performance and ensuring fair representation of minority classes. A key contribution of this study lies in its dual emphasis on performance and interpretability. To enhance interpretability, we applied SHAP to the LR model trained on ANOVA-filtered features, which provides a global understanding of the model's predictions by quantifying the impact of each selected feature on depression outcomes. SHAP values indicate the extent to which each feature contributes to the prediction, with positive values increasing the likelihood of depression and negative values decreasing it. Features are ranked by their mean absolute SHAP values, reflecting their overall importance in influencing model predictions. In our analysis, ANXI emerged as the strongest predictor of depression, followed by other relevant features such as DEPRI, POSSAT, and INFER, providing interpretable and actionable insights into the most impactful factors driving depressive risk.

By leveraging predictive modeling, statistical analysis, and interpretability, we not only demonstrate that ML can reliably detect depression but also provide actionable insights into which factors are most impactful, thereby supporting more targeted, data-driven interventions.

Ultimately, this research advances the field by presenting a comprehensive, generalizable, and methodologically sound framework for depression prediction. It contributes to both the theoretical understanding of depression risk factors and the practical development of predictive tools, offering meaningful implications for mental health professionals, policymakers, and researchers alike. Ethical integration was treated as an immediate engineering constraint for this work. For Privacy, the risk of re-identification from public data was mitigated through salted one-way hashing of identifiers and the application of data generalization techniques on semantic profiles. These measures ensured anonymization extends beyond simple direct identifier removal. Regarding Bias, disaggregated analysis revealed a significantly elevated False Negative Rate (FNR) in minority subgroups. We restored Equal Opportunity through a post-processing decision threshold re-weighting specific to those subgroups, correcting the calibration deficiency. Finally, to mitigate Operational Harm, FNR analysis demonstrated the critical asymmetry of error costs (with FNR ranging from 7.79% to 27.27%). The optimization objective was thus set to maximize Recall, reflecting the priority of minimizing FNs. The final model (LR + ANOVA) validates this strategy. The algorithm's use is strictly limited to risk stratification, mandating mandatory expert human validation.

6 Conclusion

To evaluate the robustness and relevance of this research, it is essential to contextualize its findings within the existing literature. Prior studies in depression prediction have often been constrained by limited demographic scopes, typically focusing on specific subpopulations, such as individuals within a particular age range, occupational category, or health condition [15]. While these studies have provided valuable insights, they frequently lack generalizability due to their narrow focus. Moreover, many of them primarily aim to detect depression without offering a comprehensive understanding of the contributing factors. In contrast, the present study significantly broadens the scope by incorporating a heterogeneous dataset encompassing individuals from diverse age groups, professions, and socioeconomic backgrounds. This inclusive approach enhances the generalizability and real-world applicability of our findings. Not only did we focus on predicting depression with high ACC, but we also identified and ranked the most influential predictors through rigorous feature selection techniques. Notably, the use of ANOVA for feature selection proved to be both effective and interpretable, reinforcing its utility in mental health research, especially in identifying variables with statistically significant influence on depressive outcomes.

Our best-performing ML models achieved consistently high ACC and precision scores, underscoring their potential as effective and scalable screening tools in clinical and public health settings. These models could be particularly valuable in resource-limited environments, enabling early detection and intervention through automated assessments. Furthermore, the integration of SMOTE to address class imbalance represents a methodological strength, improving model performance and ensuring fair representation of minority classes. A critical methodological innovation in this study was the use of SHAP for explainability and interpretability. SHAP quantifies the contribution of each feature to the model's output, regardless of the underlying algorithm. In this experiment, SHAP was applied to interpret the model's predictions and understand the relative importance of features. Our results indicate that SHAP provides clear insights into feature effects, offering a transparent explanation of the model's behavior. This suggests that SHAP is a powerful tool for making ML models interpretable. In future work, we plan to extend SHAP-based analysis to deep learning models and to include explainability-driven analysis of image data.

References

[1] "WHO EMRO | What you can do | Mental health." Accessed: Mar. 13, 2024. [Online]. Available: https://www.emro.who.int/mnh/what-you-can-do/index.html#accordionpan4

[2] C. Otte et al., "Major depressive disorder," Nat. Rev. Dis. Primer, vol. 2, no. 1, Art. no. 1, Sep. 2016, doi: 10.1038/nrdp.2016.65.

[3] Kolenik T, Schiepek G, Gams M. Computational Psychotherapy System for Mental Health Prediction and Behavior Change with a Conversational Agent. *Neuropsychiatr Dis Treat*. 2024;20:2465-2498 https://doi.org/10.2147/NDT.S417695

[4] A. Thapar, S. Collishaw, D. S. Pine, and A. K. Thapar, "Depression in adolescence," The Lancet, vol. 379, no. 9820, pp. 1056–1067, Mar. 2012, doi: 10.1016/S0140-6736(11)60871-4.

[5] G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, "Machine Learning in Psychometrics and Psychological Research," Front. Psychol., vol. 10, p. 2970, Jan. 2020, doi: 10.3389/fpsyg.2019.02970.

[6] Kolenik, T. (2022). Methods in Digital Mental Health: Smartphone-Based Assessment and Intervention for Stress, Anxiety, and Depression. In: Comito, C., Forestiero, A., Zumpano, E. (eds) Integrating Artificial Intelligence and IoT for Advanced Health Informatics. Internet of Things. Springer, Cham. https://doi.org/10.1007/978-3-030-91181-2_7

[7] Kolenik T, Gams M. Intelligent Cognitive Assistants for Attitude and Behavior Change Support in Mental Health: State-of-the-Art Technical Review. Electronics. 2021; 10(11):1250. https://doi.org/10.3390/electronics10111250

[8] T. Kolenik and M. Gams, "Persuasive Technology for Mental Health: One Step Closer to (Mental Health Care) Equality?," in *IEEE Technology and Society Magazine*, vol. 40, no. 1, pp. 80-86, March 2021, doi: 10.1109/MTS.2021.3056288.

[9] Moustati, I., & Gherabi, N. (2025). Deep learning applications in the internet of behaviors: a

comprehensive cross-domain survey. EDPACS, 1–27. https://doi.org/10.1080/07366981.2025.2518821

[10] Kolenik T. Intelligent Cognitive System for Computational Psychotherapy with a Conversational Agent for Attitude and Behavior Change in Stress, Anxiety, and Depression. Informatica (Slovenia). 2025;49(2):451-454. doi:10.31449/inf.v49i2.8738

[11] H. El Massari, N. Gherabi, S. Mhammedi, H. Ghandi, F. Qanouni, and M. Bahaj, "Integration of ontology with machine learning to predict the presence of covid-19 based on symptoms," Bull. Electr. Eng. Inform., vol. 11, no. 5, pp. 2805–2816, Oct. 2022, doi: 10.11591/eei.v11i5.4392.

[12] A. A. Aouragh, M. Bahaj, and N. Gherabi, "Comparative Study of Dimensionality Reduction Techniques and Machine Learning Algorithms for Alzheimer's Disease Classification and Prediction," in 2022 IEEE 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Dec. 2022, pp. 1–6. doi: 10.1109/ICECOCS55148.2022.9983211.

[13] H. E. Massari, N. Gherabi, S. Mhammedi, Z. Sabouri, H. Ghandi, and F. Qanouni, "Effectiveness of applying Machine Learning techniques and Ontologies in Breast Cancer detection," Procedia Comput. Sci., vol. 218, pp. 2392–2400, Jan. 2023, doi: 10.1016/j.procs.2023.01.214.M

[14] J. Cvetković, "Breast Cancer Patients' Depression Prediction by Machine Learning Approach," Cancer Invest., vol. 35, no. 8, pp. 569–572, Sep. 2017, doi: 10.1080/07357907.2017.1363892.M

[15] A. Grzenda et al., "Machine Learning Prediction of Treatment Outcome in Late-Life Depression," Front. Psychiatry, vol. 12, 2021, Accessed: Feb. 15, 2024. [Online]. Available: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2021.738494

[16] Z. Sabouri, N. Gherabi, M. Nasri, A. Mohamed, H. el Massari, and I. Moustati, "Prediction of Depression via Supervised Learning Models: Performance Comparison and Analysis," Int. J. Online Biomed. Eng. IJOE, vol. 19, pp. 93–107, Jul. 2023, doi: 10.3991/ijoe.v19i09.39823.

[17] K.-S. Na, S.-E. Cho, Z. W. Geem, and Y.-K. Kim, "Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm," Neurosci. Lett., vol. 721, p. 134804, Mar. 2020, doi: 10.1016/j.neulet.2020.134804.

[18] P. M. D. R. Vincent, N. Mahendran, J. Nebhen, N. Deepa, K. Srinivasan, and Y.-C. Hu, "Performance Assessment of Certain Machine Learning Models for Predicting the Major Depressive Disorder among IT Professionals during Pandemic times," Comput. Intell. Neurosci., vol. 2021, p. e9950332, Apr. 2021, doi: 10.1155/2021/9950332.

[19] Z. Jan et al., "The Role of Machine Learning in Diagnosing Bipolar Disorder: Scoping Review," J. Med. Internet Res., vol. 23, no. 11, p. e29749, Nov. 2021, doi: 10.2196/29749.

[20] Md. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, "An in-depth analysis of machine learning approaches to predict depression," Curr. Res. Behav. Sci., vol. 2, p. 100044, Nov. 2021, doi: 10.1016/j.crbeha.2021.100044.

[21] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, "Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence," Sci. Rep., vol. 11, no. 1, p. 1980, Jan. 2021, doi: 10.1038/s41598-021-81368-4.

[22] C. M. Hatton, L. W. Paton, D. McMillan, J. Cussens, S. Gilbody, and P. A. Tiffin, "Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare," J. Affect. Disord., vol. 246, pp. 857–860, Mar. 2019, doi: 10.1016/j.jad.2018.12.095.

[23] S. Natarajan, A. Prabhakar, N. Ramanan, A. Bagilone, K. Siek, and K. Connelly, "Boosting for Postpartum Depression Prediction," in 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Jul. 2017, pp. 232–240. doi: 10.1109/CHASE.2017.82.

[24] S. Jiménez-Serrano, S. Tortajada, and J. M. García-Gómez, "A Mobile Health Application to Predict Postpartum Depression Based on Machine Learning," Telemed. E-Health, vol. 21, no. 7, pp. 567–574, Jul. 2015, doi: 10.1089/tmj.2014.0113.

[25] M. S. Zulfiker, "Sabab31/Depression-Repository." Jan. 10, 2021. Accessed: Feb. 15, 2024. [Online]. Available: https://github.com/Sabab31/Depression-Repository.

[26] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," arXiv preprint arXiv:1802.03888v3, 2019. doi: 10.48550/arXiv.1802.03888.

[27] A. A. Soladoye, N. Aderinto, D. Osho, and D. B. Olawade, "Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data," *International Journal of Medical Informatics,* vol. 204, p. 106093, 2025. doi: 10.1016/j.ijmedinf.2025.106093.