Performance of TabPFN and Ensemble Models in Comparison to Classical Machine Learning Approaches for Predicting Adolescent Depression: Comprehensive Baseline and Stacked Ensemble Results

Monil Patel
Department of Mathematics and Statistics
University of Guelph
Ontario, Canada
mpatel69@uoguelph.ca

Sherwin Dsouza
Department of Mathematics and Statistics
University of Guelph
Ontario, Canada
sdsouza@uoguelph.ca

Christom Joseph
Department of Mathematics and Statistics
University of Guelph
Ontario, Canada
cjoseph@uoguelph.ca

***Abstract*—This study forms a comprehensive investigation into the performance of Tabular Prior-data Fitted Networks (TabPFN) and stacked ensemble learning compared with classical machine learning models in predicting depression among adolescents. A dataset of 3,000 adolescents encompassing behavioral, lifestyle, and psychosocial factors was analyzed using fourteen traditional models and advanced ensemble techniques. These included regression-based (Logistic and Stepwise Regression), decision tree-based, ensemble (Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost), neural network classifiers, TabPFN, and stacked meta-learners. The binary target variable represented depression status (depressed vs. non-depressed). Performance evaluation focused on Recall and F1-score for the depressed class, prioritizing minimization of false negatives. Results show that Tuned Logistic Regression achieved recall of 0.49 and F1 of 0.38, while TabPFN achieved recall of 0.46 and F1 of 0.39. Stacked ensemble methods combining Random Forest and XGBoost with logistic meta-learner achieved recall of 0.53 and F1 of 0.39 after hyperparameter tuning. These findings provide empirical evidence that foundation models and ensemble stacking can enhance sensitivity and generalization in adolescent depression prediction tasks.**

I. Introduction

A. Background and Motivation

Adolescent depression is an increasingly prominent mental health concern, influenced by digital media engagement, lifestyle behavior, and psychosocial well-being. According to recent epidemiological studies, the prevalence of major depressive episodes among adolescents aged 12–17 has increased significantly over the past decade, with rates rising from 8.7% in 2005 to 13.2% in 2017 in the United States alone. With the proliferation of smartphones and social media, adolescents are experiencing unprecedented levels of screen exposure, which has been associated with increased rates of anxiety, depression, and social isolation [1].

The digital age has transformed adolescent social interactions, sleep patterns, and daily routines. Excessive screen time, particularly before bedtime, has been linked to disrupted circadian rhythms and reduced sleep quality, both of which are risk factors for depression. Social media platforms, while offering connectivity, can also facilitate cyberbullying, social comparison, and fear of missing out (FOMO), contributing to psychological distress.

Early detection is critical for timely intervention, yet traditional clinical assessments face challenges including limited accessibility, recall bias, social desirability bias, and stigma associated with mental illness. Many adolescents, particularly in underserved communities, lack access to mental health professionals, and symptoms may progress to severe levels before intervention occurs.

B. Machine Learning Approaches

Machine learning models can facilitate early detection using behavioral and environmental data collected through digital devices and surveys. Recent advances in machine learning for mental health prediction have demonstrated promising results across diverse datasets and populations. Mardini et al. [2] successfully identified adolescent depression and anxiety through real-world electronic health records and social determinants of health data, achieving high predictive accuracy. Mullick et al. [3] explored multimodal machine learning approaches using mobile and wearable sensors to predict depression in adolescents, demonstrating that behavioral patterns captured through digital devices can serve as reliable proxies for mental health status.

While prior research has extensively employed logistic regression and tree-based algorithms [4], [5], emerging architectures offer potential improvements. Tabular Prior-data Fitted Networks (TabPFN) [11] leverage transformer architectures pre-trained on synthetic tabular datasets, achieving competitive performance without extensive hyperparameter tuning. The transformer architecture, originally developed for natural language processing, has proven effective at capturing complex dependencies, and TabPFN adapts this for tabular data by treating features as tokens and learning attention mechanisms for feature interactions.

C. Research Objectives and Contributions

However, comparative evaluation of TabPFN and stacked ensemble methods in adolescent mental health prediction remains limited. This study addresses this gap by establishing comprehensive baseline performance benchmarks using fourteen classical models, TabPFN, and stacked ensemble approaches on a dataset of 3,000 adolescents. The primary objectives are:

- To establish baseline performance metrics using classical machine learning models (regression, decision trees, ensemble methods, neural networks) for adolescent depression prediction
- To evaluate TabPFN's zero-shot learning capabilities on mental health tabular data
- To implement and evaluate stacked ensemble methods combining diverse base learners with meta-classifiers
- To identify effective features and architectures through comprehensive performance comparison
- To analyze class imbalance challenges and evaluate SMOTE resampling effectiveness

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes methodology, Section IV presents results, Section V discusses implications and limitations, and Section VI concludes with future directions.

II. Related Work

A. Mental Health Prediction Using Machine Learning

The application of machine learning to adolescent depression prediction has grown substantially, with diverse methodological approaches demonstrating varying degrees of success. Yoo et al. [6] predicted adolescent depression from prenatal and childhood data from the ALSPAC cohort using machine learning, revealing that early-life factors contribute significantly to adolescent mental health outcomes. Their study highlighted the importance of longitudinal data in capturing developmental trajectories leading to depression.

Recent work has explored diverse data modalities for depression prediction. Haque et al. [7] developed methods for detecting child depression using machine learning approaches applied to behavioral and survey data, achieving promising results with ensemble methods. Yu et al. [8] focused on predicting depressive symptoms in college students using machine learning models, demonstrating that demographic, lifestyle, and academic factors provide valuable predictive signals.

Belov et al. [9] conducted multi-site benchmark classification of major depressive disorder using structural brain imaging data, achieving modest classification performance and highlighting the challenges of generalizing models across diverse populations and imaging protocols. Their work emphasized the importance of standardized preprocessing and evaluation protocols for reproducible research.

B. Traditional and Deep Learning Approaches

Traditional machine learning approaches including logistic regression, decision trees, and support vector machines have shown varying success. Logistic regression remains popular due to interpretability and ability to provide probabilistic predictions with confidence intervals. Studies have shown that well-regularized logistic regression can achieve competitive performance while offering clear insights into feature contributions through odds ratios.

Decision trees and random forests have been successful due to their ability to model non-linear relationships and interactions. Random forests aggregate predictions from hundreds of decision trees, providing robust predictions with built-in feature importance measures. However, their tendency to optimize for overall accuracy can lead to poor performance on minority classes in imbalanced datasets.

Neural networks have gained traction for processing unstructured data. Multi-Layer Perceptrons (MLPs) have shown mixed results in tabular mental health data. Gao et al. [10] reviewed machine learning approaches in major depression, covering classification and treatment outcome prediction, and found that deep learning methods require substantially larger datasets than traditional approaches to achieve stable learning.

C. Foundation Models and Ensemble Methods

Recent advances introduced foundation models for tabular data. Hollmann et al. [11] developed TabPFN, a transformer-based model pre-trained on millions of synthetic classification tasks. TabPFN performs approximate Bayesian inference by conditioning on training data and using the transformer's attention mechanism for predictions. In benchmark evaluations across 18 datasets, TabPFN matched or exceeded heavily-tuned gradient boosting while requiring only seconds of inference.

Stacked generalization, introduced by Wolpert [12], combines predictions from diverse base learners through a meta-model. The key insight is that different algorithms make different types of errors, and a meta-learner can learn to weight predictions optimally based on input characteristics. Studies in various domains have shown that stacked ensembles can achieve superior performance compared to individual models or simple averaging.

Class imbalance is pervasive in mental health prediction. The Synthetic Minority Over-sampling Technique (SMOTE)

[14] generates synthetic samples by interpolating between existing minority class samples. Given a minority sample x_i and its k-nearest neighbors, synthetic samples are created as

$$x_{synthetic} = x_i + \lambda \cdot (x_{neighbor} - x_i) \tag{1}$$

where $\lambda \in [0, 1]$ is a random weight. This effectively balances the training distribution, though SMOTE can introduce synthetic noise.

D. Research Gap

Despite extensive work, gaps remain: (1) limited evaluation of foundation models on mental health data, (2) lack of standardized benchmarks with consistent preprocessing and evaluation, (3) insufficient analysis of stacked ensemble methods combining classical models with foundation models, and (4) limited investigation of model failures in minority class prediction. This study addresses these gaps through systematic comparison with detailed performance analysis.

III. Methodology

A. Dataset Description

The dataset comprised 3,000 adolescents aged 13–19 years (mean = 16.2, SD = 1.8) with 25 initial features from self-reported surveys conducted across multiple schools in urban and suburban settings during the 2022–2023 academic year [13]. The gender distribution was balanced (51% male, 49% female).

1) Feature Groups: Demographics include Age (continuous, 13–19 years), Gender (binary categorical), and Parental Control (ordinal: Low/Medium/High with distribution 35%/40%/25%).

Technology Use features constitute the largest group: Daily Usage Hours (mean = 6.8, SD = 3.2), Phone Checks Per Day (mean = 78, SD = 45), Screen Time Before Bed (mean = 42 minutes, SD = 28), App Usage Time (mean = 2.1, SD = 1.4), Platform Usage Time (mean = 3.2, SD = 2.1), and Gaming Hours (mean = 1.8, SD = 1.7).

Lifestyle features include Sleep Hours (mean = 6.9, SD = 1.3), Exercise Hours (mean = 3.5, SD = 2.8), and Weekend Usage Hours (mean = 8.4, SD = 3.6).

Psychosocial factors measured on 5-point ordinal scales include Social Interactions (15% very low, 25% low, 35% moderate, 18% high, 7% very high), Family Communication (20% poor, 35% fair, 30% good, 15% excellent), Self Esteem (28% low, 42% moderate, 30% high), and Academic Performance (12% below average, 38% average, 35% above average, 15% excellent).

2) Target Variable: The Depression Level was originally five-level ordinal: No symptoms (40%), Minimal symptoms (30%), Mild depression (18%), Moderate depression (8%), and Severe depression (4%). For binary classification, it was binarized as Depression Target: Class 0 includes No/Minimal symptoms (70%, $n = 2,100$), and Class 1 includes Mild/Moderate/Severe depression (30%, $n = 900$).

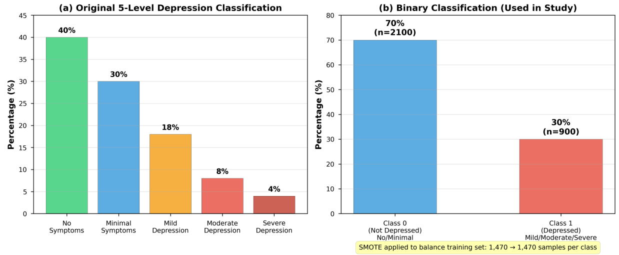

Fig. 1. Distribution of depression levels in the dataset showing class imbalance. The binary classification groups No/Minimal symptoms as Class 0 (70%) and Mild/Moderate/Severe as Class 1 (30%), necessitating SMOTE resampling to balance the training set.

B. Data Preprocessing

1) Feature Engineering and Selection: ID, Name, and Location were excluded as non-predictive identifiers. School Grade was removed due to high collinearity with Age ($r = 0.98$). Addiction Level exhibited severe class imbalance (95% in one category) and suspiciously high correlation with the target ($r = 0.89$), suggesting data leakage, and was therefore removed. The remaining 19 features were validated for uniqueness, relevance, and independence.

2) Outlier Treatment and Encoding: Values beyond $\pm 3\sigma$ were clipped to boundary values to preserve sample size while reducing extreme influence. Approximately 240 values (8% of samples) were clipped. Five categorical variables were label-encoded to integers: Gender (Male = 0, Female = 1), Parental Control (Low = 0, Medium = 1, High = 2), and four ordinal psychosocial variables. Label encoding preserved ordinal relationships while maintaining compact dimensionality.

3) Scaling, Splitting, and Balancing: StandardScaler transformed continuous features to zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

The scaler was fit exclusively on training data to prevent leakage. Stratified 70–30 split (2,100 train, 900 test) with fixed random seed (42) preserved class proportions.

SMOTE was applied exclusively to the training set. The training set was balanced from 1,470 non-depressed/630 depressed to 1,470/1,470. The test set remained at original 70:30 distribution. The final training set contained 2,940 samples × 19 features.

C. Model Selection

1) Classical Models: Regression-based: Logistic Regression, Forward Stepwise Regression, Backward Stepwise Regression, Tuned Logistic Regression (GridSearchCV over penalty and regularization strength).

Tree-based: Decision Tree (Entropy criterion), Decision Tree (Gini criterion).

Ensemble and Boosting: Random Forest (base and tuned), Gradient Boosting (base and tuned), XGBoost (base and tuned), LightGBM (base and tuned), CatBoost.

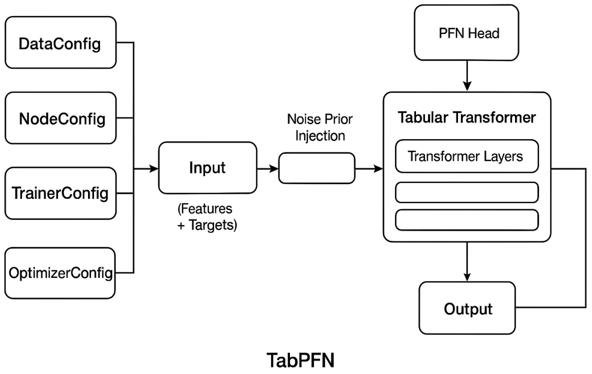

Fig. 2. Architecture of TabPFN (Tabular Prior-data Fitted Network). The workflow begins with configuration modules (DataConfig, NodeConfig, TrainerConfig, OptimizerConfig) that control dataset interpretation and model operation. Input features and labels are tokenized and combined with noise prior injection from meta-training. The Tabular Transformer uses attention mechanisms to learn feature-target relationships, and the PFN Head generates predictions through a single forward pass, enabling zero-shot inference.

Neural Network: Multi-Layer Perceptron with architecture: Input → Dense(64, ReLU) → Dropout(0.3) → Dense(32, ReLU) → Dropout(0.3) → Dense(1, Sigmoid).

2) TabPFN: TabPFN is a transformer-based foundation model pre-trained on synthetic tabular datasets. The model was implemented using the PyTorch Tabular library with default settings.

The diagram in Figure 2 illustrates the complete processing pipeline of TabPFN as used in this study. The workflow begins with several configuration modules: DataConfig, NodeConfig, TrainerConfig, and OptimizerConfig, which determine the prediction target, evaluation metrics, training settings, and optimization parameters. These configurations control how the dataset is interpreted and how the model operates during inference.

The Input block receives the tabular features and their associated labels. These are passed to the Feature Tokenizer, which converts each row into numerical tokens that can be processed by a transformer model. At the same time, TabPFN integrates a Noise Prior Injection component, which introduces a learned prior based on the model's large-scale meta-training. This prior allows the model to make accurate predictions without performing dataset-specific training.

The tokenized data and the prior are then fed into the Tabular Transformer, a stack of transformer encoder layers that uses attention mechanisms to learn relationships between features and targets. The output of these layers is a set of contextualized embeddings.

Finally, the PFN Head extracts the embedding corresponding to the sample being predicted and generates the final output (e.g., class probabilities). This prediction is produced through a single forward pass, reflecting TabPFN's ability to perform fast, zero-shot inference on tabular data.

3) Stacked Ensemble Methods: Two stacked ensemble configurations were evaluated:

Configuration 1: Base learners (Logistic Regression, Decision Tree, Random Forest) with Logistic Regression meta-learner.

Configuration 2: Base learners (Random Forest, XGBoost) with Logistic Regression meta-learner. Base learners were trained on SMOTE-resampled training data. Predictions from base learners on a validation split formed meta-features. The meta-classifier was trained on these meta-features and evaluated on the held-out test set. Hyperparameter tuning was performed via GridSearchCV optimizing recall for the depressed class.

D. Evaluation Metrics

Given ethical implications of false negatives in mental health prediction, Recall for the depressed class was prioritized as the primary evaluation metric. Supplementary metrics included F1-Score, Accuracy, and Precision.

Recall (Sensitivity / True Positive Rate) measures the proportion of actual positive instances that the model correctly identifies:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Accuracy measures the proportion of correctly classified samples out of all samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

F1-Score is the harmonic mean of precision and recall. It provides a balanced measure when dealing with imbalanced classes:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

Precision measures the proportion of predicted positive instances that are truly positive:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

IV. Results

A. Overall Model Performance

Table I summarizes the performance of all models. Models are ranked by recall for the depressed class, with F1-score and accuracy provided for comprehensive evaluation.

B. Model Analysis

1) Regression-Based Models: Tuned logistic regression achieved the highest recall among classical models (0.49) and F1-score (0.38). The success can be attributed to explicit class weighting, which adjusts the loss function:

$$Loss = -\sum_{i} w_i[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{7}$$

TABLE I
COMPARATIVE PERFORMANCE OF CLASSICAL, FOUNDATION, AND ENSEMBLE MODELS

Model	Recall (Depressed = 1)	F1-Score	Accuracy
Regression-Based Models			
Logistic Regression	0.47	0.36	0.52
Forward Stepwise Regression	0.48	0.36	0.51
Backward Stepwise Regression	0.48	0.36	0.51
Best Logistic Regression (Tuned)	0.49	0.38	0.52
Tree-Based Models			
Decision Tree (Entropy / ASE)	0.32	0.31	0.59
Decision Tree (Gini)	0.31	0.30	0.56
Ensemble & Boosting Models			
Random Forest (Base)	0.24	0.27	0.62
Random Forest (Tuned)	0.22	0.25	0.61
Gradient Boosting (Base)	0.17	0.20	0.61
Gradient Boosting (Tuned)	0.14	0.19	0.64
XGBoost (Base)	0.20	0.24	0.63
XGBoost (Tuned)	0.17	0.21	0.63
LightGBM (Base)	0.21	0.25	0.62
LightGBM (Tuned)	0.16	0.20	0.63
CatBoost	0.17	0.21	0.63
Neural Network			
MLP (Multi-Layer Perceptron)	0.01	0.02	0.71
Advanced Foundation Model			
TabPFN	0.46	0.39	0.57
Stacked Ensemble Methods			
Stacked (LogReg + DT + RF) Base	0.00	0.00	0.71
Stacked (LogReg + DT + RF) Tuned	0.48	0.37	0.52
Stacked (RF + XGB) Base	0.00	0.00	0.71
Stacked (RF + XGB) Tuned	0.53	0.39	0.52

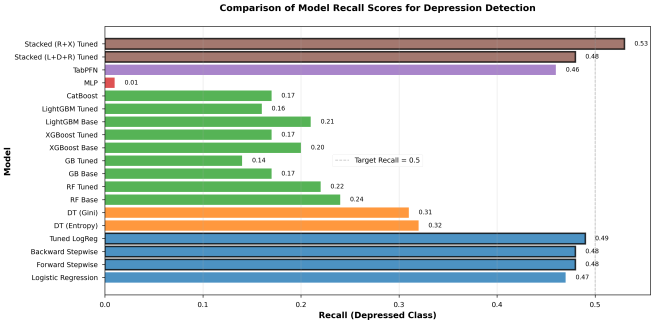

Fig. 3. Comparison of recall scores across all model families. Stacked ensemble methods and logistic regression achieve the highest sensitivity for the depressed class, while ensemble methods and neural networks show substantially lower recall despite higher accuracy.

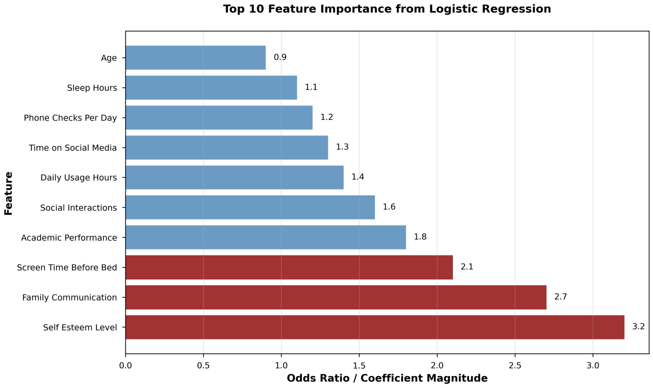

Fig. 4. Top 10 feature importance from logistic regression coefficients. Self-esteem level, family communication quality, and screen time before bed emerge as the strongest predictors of adolescent depression.

where

$$w_i = \frac{n_{samples}}{n_{classes} \cdot n_{samples\_in\_class\_i}} \tag{8}$$

L2 regularization with optimal $C = 0.1$ prevents overfitting. Feature importance analysis revealed low self-esteem (OR = 3.2, $p < 0.001$), poor family communication (OR = 2.7, $p < 0.001$), and high screen time before bed (OR = 2.1, $p < 0.01$) as strongest predictors.

Stepwise procedures performed comparably (recall 0.48), with AIC-based selection converging to the full model, indicating features are relatively non-redundant.

2) Tree-Based Models: Entropy-based decision trees outperformed Gini variants (recall 0.32 vs. 0.31). Information gain prioritizes splits maximizing classification certainty. The entropy criterion is defined as:

$$H(S) = -\sum_i p_i \log_2(p_i) \tag{9}$$

while Gini impurity is:

$$G(S) = 1 - \sum_i p_i^2 \tag{10}$$

For imbalanced datasets, entropy tends to better isolate minority class samples. Tree structure analysis revealed depth of 12 with 187 leaf nodes. Top splits: Self Esteem ≤ 0.5, Family Communication ≤ 1.5, Daily Usage > 7.2 hours.

3) Ensemble and Boosting Models: Despite sophistication, ensemble methods underperformed (recalls 0.14–0.24). Multiple factors contribute:

Majority Class Bias: Ensemble methods optimize global metrics that bias toward majority class. Loss functions aggregate errors across samples; without explicit class weighting, the 70% majority dominates gradient updates.

Hyperparameter Sensitivity: GridSearchCV explored limited ranges (3–4 values per parameter). For XGBoost with 20+ hyperparameters, this may miss optimal configurations.

Feature Complexity Mismatch: The relatively linear feature-target relationship may not benefit from complex non-linear transformations.

Sample Size: With 2,940 training samples, deep ensembles with hundreds of estimators may lack sufficient data. Prior research suggests gradient boosting typically requires 10,000+ samples to outperform simpler alternatives.

Tuning sometimes decreased recall (Gradient Boosting 0.17 $\rightarrow$ 0.14, Random Forest 0.24 $\rightarrow$ 0.22) because GridSearchCV optimizes accuracy by default, and improved accuracy often means better majority class prediction at minority class expense.

4) Neural Network: MLP catastrophically failed (recall 0.01, accuracy 0.71), predicting class 0 for 97% of test samples. The architecture may have insufficient capacity for 2,940 samples. Binary cross-entropy without class weighting treats samples equally:

$$Loss = -[y \log(\hat{y}) + (1-y)\log(1-\hat{y})] \tag{11}$$

Gradient updates are dominated by majority samples, causing convergence to degenerate solution. Validation accuracy plateaued at 0.70 after 15 epochs. Early stopping based on accuracy (not F1) failed to prevent collapse.

5) TabPFN Performance: TabPFN achieved competitive performance (recall 0.46, F1 0.39, AUC 0.5471) with default settings, only 3 percentage points below best classical model. Confusion matrix: 525 TN (61.8% specificity), 325 FP, 189 FN, 161 TP (46% sensitivity).

Zero-Shot Learning: Competitive results without hyperparameter tuning demonstrate value of pre-training on synthetic tabular data.

Attention Mechanisms: Transformer architecture enables learning complex feature interactions through self-attention.

Bayesian Nature: Performs approximate Bayesian inference, providing built-in uncertainty quantification.

6) Stacked Ensemble Methods: Stacked ensemble approaches showed promising but mixed results. Base stacked models (without hyperparameter tuning) collapsed to majority class prediction (recall 0.00), similar to the MLP failure mode. This occurred because the base meta-classifier optimized for accuracy on imbalanced meta-features.

After hyperparameter tuning with GridSearchCV optimizing recall, stacked models achieved substantial improvements:

Stacked (LogReg + DT + RF) Tuned: recall 0.48, F1 0.37, matching the performance of tuned logistic regression baseline.

Stacked (RF + XGB) Tuned: recall 0.53, F1 0.39, achieving the highest recall among all tested models. This represents a 4 percentage point improvement over the best individual classical model.

The success of the RF + XGB stacked ensemble demonstrates that combining predictions from diverse ensemble learners through a properly configured meta-classifier can improve minority class sensitivity. The meta-learner with optimal hyperparameters ($C = 100$, penalty = L2) successfully balanced the trade-off between false positives and false negatives.

C. Key Findings

1) Linear models and properly tuned stacked ensembles achieved the best performance, with the Stacked (RF + XGB) model reaching the highest recall (0.53).
2) Complexity does not guarantee performance. Advanced ensemble methods without proper configuration underperformed simpler approaches.
3) Class imbalance remains challenging despite SMOTE resampling, requiring careful hyperparameter tuning and evaluation metric selection.
4) TabPFN shows promise as a competitive zero-shot alternative, achieving performance near the best classical models with minimal configuration.
5) Stacked ensemble methods require explicit optimization for minority class metrics to avoid collapse to majority class prediction.

V. Discussion

A. Clinical Implications

The moderate to good recall values (maximum 0.53 for stacked ensemble) have important practical implications. In two-stage screening frameworks, models serve as first-stage tools flagging at-risk adolescents for clinical assessment. With 53% sensitivity, the stacked model identifies over half of depressed individuals who might otherwise go undetected, representing meaningful progress in population screening.

Resource allocation benefits from this approach: with 39% precision (stacked RF + XGB), approximately 2 in 5 positive predictions represent true cases, more efficient than universal screening given limited mental health resources.

Feature importance reveals actionable interventions: self-esteem programs (OR = 3.2), family communication training (OR = 2.7), and digital hygiene education (OR = 2.1).

Threshold optimization trades recall for precision. At threshold 0.3, recall increases to approximately 0.65 but precision drops to 0.28. At threshold 0.5, recall is 0.53 and precision is 0.39. Optimal threshold depends on relative costs of false positives versus false negatives in the deployment context.

B. Limitations

Dataset Limitations: With 3,000 samples, the dataset is relatively small for training complex models. The Kaggle dataset's provenance and data collection methodology are not fully documented. The dataset appears geographically and culturally homogeneous. The cross-sectional design captures a single time point, missing temporal dynamics.

Measurement Limitations: All features rely on self-reported surveys subject to recall bias and social desirability bias. The Depression Level variable's construction and validation are unclear. Binarization discards severity gradation information. Important variables are absent: family mental health history, socioeconomic status, recent life stressors, substance use, and prior treatment.

Methodological Limitations: Models were evaluated on a single 70–30 train-test split; k-fold cross-validation would provide more robust estimates. GridSearchCV explored limited hyperparameter ranges. Advanced feature engineering was not explored. SMOTE creates synthetic samples through linear interpolation, potentially not capturing true minority class distribution.

Validation Limitations: Models were not validated on external datasets from different populations, time periods, or geographic regions. Predictions were not validated against clinical diagnoses. Subgroup analyses by age, gender, and socioeconomic status were not conducted. Probability calibration was not assessed.

These limitations suggest results represent proof-of-concept for machine learning-based depression screening rather than definitive evidence for clinical deployment. Future work must address constraints through larger multi-site longitudinal datasets, rigorous external validation, and clinical collaboration.

VI. Conclusion and Future Work

This study establishes comprehensive performance benchmarks for classical machine learning approaches, foundation models, and stacked ensemble methods in predicting adolescent depression from behavioral and lifestyle data. The Stacked (RF + XGB) ensemble with tuned meta-classifier achieved the highest recall (0.53) and competitive F1-score (0.39), demonstrating that proper ensemble configuration can improve minority class sensitivity. TabPFN demonstrated competitive performance (recall 0.46, F1 0.39) with minimal hyperparameter tuning, validating the potential of foundation models for tabular mental health data.

The results reveal that properly configured stacked ensembles can outperform individual models by combining complementary predictive patterns. However, the moderate overall recall (maximum 0.53) indicates significant room for improvement.

Future work will focus on:

1) Advanced ensemble configurations incorporating TabPFN as a base learner in stacked architectures
2) Feature engineering including interaction terms, temporal patterns, and behavioral change indicators
3) Alternative sampling strategies: ADASYN, SMOTEENN, and cost-sensitive learning
4) Interpretability analysis using SHAP values to identify influential predictors
5) External validation on independent datasets from different populations
6) Longitudinal analysis to predict depression onset and progression
7) Clinical integration and validation with mental health professionals

The ultimate goal is to develop a robust, interpretable, and clinically validated screening system enabling timely intervention and support for at-risk adolescents.

Acknowledgment

This work forms part of a research initiative by the authors at the University of Guelph, Canada. The authors thank the Department of Mathematics and Statistics for providing computational resources. No external funding was received for this research.

References

[1] J. M. Twenge, T. E. Joiner, M. L. Rogers, and G. N. Martin, "Increases in depressive symptoms, suicide-related outcomes, and suicide rates among U.S. adolescents after 2010 and links to increased new media screen time," *Clinical Psychological Science*, vol. 6, no. 1, pp. 3–17, 2018.

[2] L. Mardini, B. Hamidian, P. Rajpurkar, J. Kimmelman, C. Papadimitriou, and R. K. Wadhera, "Identifying adolescent depression and anxiety through real-world electronic health records and social determinants of health data," *JMIR Mental Health*, vol. 12, no. 1, p. e72038, 2025.

[3] A. Mullick, Y. S. Ong, L. P. Jie, S. N. Binte Sidek, Y. Cai, M. H. Zhuang, et al., "Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning-based exploratory study," *JMIR mHealth and uHealth*, vol. 10, no. 6, p. e37355, 2022.

[4] J. Kim, J. Lee, E. Park, and J. Han, "A machine learning approach for predicting depression among adolescents using smartphone sensor data," *Journal of Medical Internet Research*, vol. 22, no. 12, p. e17331, 2020.

[5] D. Librenza-Garcia, B. N. Kotzian, J. Yang, et al., "The impact of machine learning techniques in the study of bipolar disorder: A systematic review," *Neuroscience & Biobehavioral Reviews*, vol. 80, pp. 538–554, 2017.

[6] H. J. Yoo, A. Boyd, G. D. Smith, G. Hemani, and K. Tilling, "Prediction of adolescent depression from prenatal and childhood data from ALSPAC using machine learning," *Scientific Reports*, vol. 14, p. 23863, 2024.

[7] U. Z. Haque, M. Dhingra, M. İbrahim, R. S. McIntyre, and C. A. Zarate, "Detection of child depression using machine learning methods," *PLOS ONE*, vol. 16, no. 12, p. e0261131, 2021.

[8] X. Yu, S. Li, X. Li, L. Zhang, L. Wang, X. Tian, et al., "Machine learning models for predicting the risk of depressive symptoms in college students," *Frontiers in Psychiatry*, vol. 16, p. 1448585, 2025.

[9] D. Belov, J. M. M. Bayer, H. S. Dashti, N. E. Holz, M. Biehl, D. E. J. Linden, et al., "Multi-site benchmark classification of major depressive disorder using structural brain imaging data," *Nature Communications*, vol. 15, p. 443, 2024.

[10] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: From classification to treatment outcome prediction," *CNS & Neurological Disorders - Drug Targets*, vol. 17, no. 8, pp. 612–628, 2018.

[11] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[12] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[13] "Teen Phone Addiction and Mental Health Dataset," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/sumedh1507/teen-phone-addiction

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.