Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 169-177. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1716

Performance Evaluation of TabPFN for Student Depression Prediction Across Varying Sample Sizes

Florentina Yuni Arini, Muhammad Kahvi Khakam Syah, Fernando Dinar Setiawan, Rafif Musyaffa Indarto, Muhammad Danil Aminuddin, Fairuz Trideas Hilmy, Ahmad Imam Mutaqin

Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia

E-mail: floyuna@mail.unnes.ac.id

Abstract

Early prediction of depression in students is a critical challenge, often hindered by the scarcity of large, labelled datasets. While supervised tabular classifiers such as Random Forest, XGBoost, and CatBoost are powerful, they typically require sufficient data and careful hyperparameter optimisation (HPO) to generalise effectively. This paper evaluates the Tabular Prior-data Fitted Network (TabPFN), a foundation model for supervised tabular learning, as a zero-shot classifier for student depression prediction. We conduct a comparative study against five robustly configured baseline classifiers (Random Forest, XGBoost, CatBoost, SVM, and Naive Bayes) across three publicly available student mental health datasets of varying sizes sourced from Kaggle: a micro-sample dataset with 101 instances, a small-sample dataset with 7,022 instances, and a moderate-sample dataset with 27,901 instances. Dataset categorization by size is defined relative to TabPFN v2.5's operational capacity of 50,000 samples rather than general machine learning conventions. Using F1-Score as the primary evaluation metric, our empirical results demonstrate a performance crossover linked to data size. On the micro-sample, imbalanced dataset, TabPFN achieved the highest F1-Score of 0.727, outperforming the best baseline (CatBoost and Random Forest, F1 = 0.667). In the ablation study, both raw and preprocessed inputs yielded identical results for TabPFN on this dataset, highlighting its capacity to handle unprocessed data without performance loss. On the small and moderate datasets, the tuned baselines were competitive or superior, with CatBoost leading on the moderate-sample dataset (F1 = 0.869). We conclude that TabPFN is an effective and efficient baseline for depression prediction tasks in data-scarce environments, providing competitive results without HPO, while traditional ensembles remain preferred for larger datasets.

Keywords: *TabPFN, Supervised Tabular Learning, Depression Prediction, Small-Sample Learning, Machine Learning*

1. Introduction

Mental health disorders represent a growing global concern, with depression ranking as the second largest contributor to global disability burden worldwide [1]. University students are a particularly vulnerable population, facing a unique confluence of academic pressure, social adjustments, and financial stress, all of which can severely impact their mental well-being and academic performance [2], [3]. The high prevalence of mental health issues among students is a well-documented phenomenon, with meta-analytic evidence indicating that approximately one in three college students reports depressive symptoms [4].

The early and accurate prediction of at-risk students is therefore a key goal for university counselling services and public health researchers alike. In pursuit of this goal, the application of machine learning (ML) to mental health prediction has become a rapidly growing field [5], [6]. Supervised learning algorithms operating on structured tabular survey data, such as Random Forest [10], Support Vector Machines (SVM) [5], and gradient-boosted ensembles like XGBoost [12] and CatBoost [20], have formed the cornerstone of these efforts, achieving promising results in classifying depression and other conditions from demographic, academic, and psychological features [3].

However, a fundamental challenge persists. Mental health data, particularly at the institutional level, is often scarce, sensitive, and difficult to collect at scale. Researchers frequently work with datasets containing only tens or hundreds of labelled instances. In such small-sample regimes, traditional classifiers may struggle to generalise, and the extensive hyperparameter optimisation (HPO) they require becomes unreliable due to

limited validation data. This creates a practical bottleneck: the domains that most need robust prediction tools are often the ones with the least data to build them.

Building on the growing recognition that small-data machine learning models hold promise for mental health applications [18], this paper provides an empirical evaluation of TabPFN v2.5 [14] as a zero-shot classifier for student depression prediction. We benchmark TabPFN against five established baseline classifiers, including CatBoost, a commonly used gradient-boosted model often absent from prior comparisons in this domain, across three student mental health datasets of deliberately varying sizes (N = 101, N = 7,022, N = 27,901). We further conduct an ablation study comparing TabPFN's performance on raw versus preprocessed input to evaluate its practical ease of use. The study aims to provide actionable, evidence-based guidance for mental health researchers on model selection as a function of available data.

2. Related Works

This section reviews the application of machine learning techniques to mental health prediction, with particular emphasis on supervised learning approaches for tabular data and the recent emergence of tabular foundation models.

2.1. Machine Learning for Mental Health Prediction

The application of machine learning to mental health is a well-established and rapidly growing field of research [2], [5]. Researchers have leveraged data mining and machine learning algorithms to identify patterns, predict risks, and classify mental health conditions from diverse data sources. A significant body of work has focused on student populations, recognising their vulnerability to stress, anxiety, and depression [2-4].

Supervised learning algorithms operating on tabular survey data have been the cornerstone of these efforts. These methods are widely applied to classify student stress and depression levels using datasets that include psychological, academic, and social factors [2], [3]. Random Forest [10] and Support Vector Machines are frequently cited for their high accuracy in classification tasks involving structured tabular features [3], [5]. Ensemble methods, especially gradient-boosted trees, are also highly popular due to their robust performance on supervised tabular data tasks. XGBoost [12] has been benchmarked extensively, showing strong predictive power in mental health monitoring tasks. CatBoost [20], another prominent gradient-boosted framework, is widely used in applied ML due to its native handling of categorical features and its ordered boosting mechanism that reduces prediction shift. However, CatBoost has been notably absent from several student mental health prediction benchmarks, a gap this study addresses by including it as a baseline.

The simplicity and interpretability of other supervised classifiers, like Naive Bayes, have also proven effective, particularly in text-based depression detection on social media [6] and in applications for stress detection [5], [13].

Despite these advances, a common limitation across these studies is the reliance on either small institutional datasets, which risk poor generalisation, or on large-scale clinical records that may not reflect the specific context of student surveys. Moreover, the standard practice of tuning each model's hyperparameters via grid or random search can be unreliable when the dataset itself is very small, as the tuning process can easily overfit to noise in the limited validation data.

2.2. Tabular Foundation Models: TabPFN

The landscape of tabular machine learning has been significantly altered by the introduction of foundation models. The Tabular Prior-data Fitted Network, or TabPFN [7], is a novel foundation model that is pre-trained once on a massive, synthetically generated dataset of over one billion tabular samples. This pre-training, which is based on generating data from a prior distribution of causal structures, allows TabPFN to learn a general, robust prior for supervised tabular classification.

The key innovation of TabPFN is how it performs inference. It treats a new small dataset as a single prompt in the style of in-context learning (ICL). The entire training set and the test set are fed into the Transformer in a single forward pass. The model then directly outputs the predictions in seconds, without requiring any gradient descent, model tuning, or hyperparameter optimisation. This process effectively approximates a complex Bayesian inference procedure. The original TabPFN was specifically designed to excel where traditional supervised tabular models falter: on small datasets with fewer than 10,000 samples. Its authors demonstrated that it can outperform tuned ensembles of traditional models in seconds [7].

The most recent version, TabPFN v2.5 [14], has further expanded these capabilities. It supports datasets with up to 50,000 data points and 2,000 features, a 20× increase in data cells compared to v2. On the industry standard TabArena benchmark [15], [16], TabPFN v2.5 is reported as the leading method, outperforming tuned tree-based models

and matching the accuracy of AutoGluon 1.4, a complex four-hour tuned ensemble. Notably, default TabPFN v2.5 is reported to have a 100% win rate against default XGBoost on small to medium-sized classification datasets (up to 10,000 data points) [14].

A key practical advantage of TabPFN, particularly relevant for researchers without deep ML expertise, is its ability to consume raw, unprocessed tabular data and still deliver competitive predictions. Unlike traditional pipelines that require careful feature engineering, encoding, and scaling, TabPFN's in-context learning mechanism handles heterogeneous feature types natively [7], [14]. This study explicitly tests this property through an ablation comparing raw and preprocessed inputs.

2.3. Research Gap

While Wang et al. [18] have surveyed small-data ML approaches for mental health forecasting and identified TabPFN among the most promising candidate models, their work is a conceptual review rather than an empirical benchmark. The TabPFN v2.5 report [14] similarly documents over 100 published use cases across diverse domains, yet applications to student mental health survey data remain limited. The student mental health prediction literature continues to rely predominantly on traditional supervised classifiers, and empirical evidence on how TabPFN compares against robustly configured baselines, including CatBoost, on this specific task is lacking.

Furthermore, TabPFN's practical advantage of requiring no preprocessing has not been empirically tested on student mental health survey data, nor has its effectiveness been systematically compared across varying dataset sizes within this domain. This paper addresses both of these gaps by providing a controlled empirical comparison that extends the theoretical recommendations of Wang et al. [18] into actionable evidence.

3. Metodhology

This section details the datasets, the target variable construction, the machine learning models evaluated, the experimental procedure, and the metrics used to compare performance.

3.1. Data Description

This study utilises three distinct supervised tabular datasets related to student mental health, all of which are publicly available community uploads on Kaggle. Dataset categorization by size is defined relative to TabPFN v2.5's operational capacity of 50,000 samples [14], rather than general machine learning conventions. Accordingly, we label Dataset 1 (n=101) as *micro-sample*, representing just 0.2% of this capacity; Dataset 2 (n=7,022) as *small-sample*, at approximately 14% of capacity; and Dataset 3 (n=27,901) as *moderate-sample*, at approximately 56% of capacity. The characteristics of these datasets are summarised in Table 1 (next page), highlighting their diversity in sample size and feature composition.

The deliberate selection of three datasets with widely varying sample sizes (N = 101, N = 7,022, N = 27,901) enables a direct investigation of the central hypothesis: that TabPFN's advantage is most pronounced in the micro-sample regime and diminishes as data availability increases.

3.2. Target Variable Construction

Depression was selected as the target variable for three reasons. First, depression constitutes the most prevalent mental health condition among university students globally and ranks as the second largest contributor to global disability [1], making its prediction particularly relevant for early intervention efforts. Second, depression was the only mental health indicator consistently available across all three datasets, enabling cross-dataset comparison under a unified prediction task. Third, binarising the target as presence versus absence of depression aligns with the practical requirements of screening applications, where the primary objective is to identify students who may benefit from further assessment rather than to quantify symptom severity [19].

Where datasets included additional mental health indicators (e.g., Anxiety and Panic Attack in Dataset 1; Anxiety Score and Stress Level in Dataset 2), these were retained as predictor features rather than target variables. In Dataset 1, moderate co-occurrence was observed among conditions, with 27.7% of students reporting two or more conditions simultaneously. This overlap means that the retained indicators carry predictive signal for depression while remaining conceptually distinct constructs.

The specific construction for each dataset is as follows:

Dataset 1 (Student Mental Health). This dataset was collected from students at the International Islamic University Malaysia (IIUM) [9]. It contains 101 instances with 10 features, including demographic variables (Gender, Age, Course, Year of Study, Marital Status, CGPA) and

Table 1. Dataset characteristics

Property	Dataset 1 (Micro)	Dataset 2 (Small)	Dataset 3 (moderate)
Name	Student Mental Health (IIUM)	Student Mental Health Assessments	Student Depression Dataset
Total Samples	101	7,022	27,901
Train / Test Split	80 / 21	5,617 / 1,405	22,320 / 5,581
Number of Features	10	9	17
Target Variable	Depression (binary)	Depression Score ≥ 3 (binary)	Depression (binary)
Class Balance	Imbalanced (67:33)	Approximately balanced	Approximately balanced
Source	Kaggle [9]	Kaggle [21]	Kaggle [22]

three binary self-reported mental health indicators (*Do you have Depression?, Do you have Anxiety?, Do you have Panic Attack?*). Each indicator is a binary Yes/No response. The target variable is the binary "Depression" column directly. The remaining indicators (Anxiety, Panic Attack) serve as predictor features. The dataset exhibits class imbalance: approximately 67% of respondents reported no depression, and 33% reported depression.

Dataset 2 (Students Mental Health Assessments). This dataset [21] contains 7,022 instances with 9 features capturing self-rated mental health scores. The original features include a Depression Score on a 1–5 ordinal scale, alongside Anxiety Score, Stress Level, Sleep Quality, and other related variables. The target was constructed by binarising the Depression Score: instances with a score of 3 or above were labelled as positive (indicating at least moderate depression), and instances below 3 were labelled as negative. The remaining scores (Anxiety Score, Stress Level, etc.) were retained as predictor features.

Dataset 3 (Student Depression Dataset). This is the largest dataset [22] with 27,901 instances and 17 features, covering demographic variables (Gender, Age, City), academic variables (Academic Pressure, Study Satisfaction, CGPA), lifestyle factors (Sleep Duration, Dietary Habits, Work/Study Hours), financial stress, and family history of mental illness. The dataset also includes a binary variable indicating whether the student has experienced suicidal thoughts, which was retained as a predictor feature given its established clinical relevance to depression. The target variable is the binary "Depression" column provided directly in the dataset.

3.3. Evaluation Metric: F1-Score as the Primary Metric

We adopt F1-Score as the primary evaluation metric across all datasets and report Accuracy, Precision, Recall, F1-Macro, and ROC-AUC as supplementary metrics. The choice of F1-Score is motivated by the nature of the datasets and the prediction task. In mental health screening, the cost of a false negative (failing to identify a student at risk) is arguably higher than the cost of a false positive (flagging a student for follow-up who is not at risk). The F1-Score, as the harmonic mean of Precision and Recall, provides a balanced measure that penalises models which achieve high accuracy by simply predicting the majority class, a particular concern for Dataset 1, which exhibits a 67:33 class imbalance.

Using Accuracy as the sole metric in such imbalanced settings can be misleading. A model that predicts "no depression" for every instance in Dataset 1 would achieve approximately 67% accuracy while being clinically useless. F1-Score directly addresses this by requiring the model to perform well on both identifying true positives and avoiding false positives.

3.4. Models Evaluation

We evaluate TabPFN v2.5 [14] as the focal model against five baseline classifiers: Random Forest [10], XGBoost [12], CatBoost [20], SVM [11], and Naive Bayes. TabPFN v2.5 is accessed via the tabpfn Python package (v6.3.2), which loads pre-trained model weights (tabpfn-v2.5-classifier-v2.5_default.ckpt, 42.9 MB) from Hugging Face Hub (Prior-Labs/tabpfn_2_5). All inference is executed locally on the runtime GPU; no cloud API is used.

TabPFN is used in its default, zero-shot configuration. No hyperparameter tuning, feature engineering, or preprocessing was applied for the primary evaluation. The model receives the training data and test data in a single forward pass and returns predictions directly. This zero-shot approach is the core proposition being tested: can a pre-trained prior match or exceed the performance of classifiers that have been specifically configured for the task?

Baseline classifiers were configured with fixed, robust hyperparameter settings informed by best practices and preliminary experimentation, rather than exhaustive automated HPO. This design

choice was deliberate. Given that Dataset 1 contains only 101 instances (80 for training, 21 for testing), automated HPO techniques such as grid search with cross-validation are unreliable: the small validation folds introduce high variance in performance estimates, and the "optimal" hyperparameters found are likely to be artefacts of the specific fold splits rather than genuinely generalisable configurations. By using fixed, well-chosen defaults, we ensure a fairer and more reproducible comparison. For CatBoost specifically, we used a learning rate of 0.1, max depth of 6, 500 iterations, and a random seed for reproducibility. Full hyperparameter settings for all baselines are documented in the experimental notebook for reproducibility.

3.5. Experimental Setup

All experiments were conducted on Google Colaboratory using a Tesla T4 GPU. This platform was chosen deliberately for two reasons: (1) to ensure full reproducibility, as any researcher can replicate the experiments without requiring specialised hardware; and (2) because TabPFN benefits from GPU acceleration for its forward-pass inference, and the T4 represents a widely accessible GPU tier [7], [14].

Table 2. Software environment. HF = hugging face hub; GC = google colab.

Component	Version	Source
tabpfn (inference package)	6.3.2	PyPI
TabPFN model weights	v2.5 (v2.5_default.ckpt)	HF
Python	3.11	GC
PyTorch	2.9.0+cu126	GC
CUDA	12.6	GC
Pandas	2.2.2	PyPI
NumPy	2.0.2	PyPI
Hardware	Tesla T4 GPU (15 GB VRAM)	GC

Table 2 lists the software versions used in this study. We note that the tabpfn Python package and the TabPFN model weights are independently versioned: the package (v6.3.2) provides the inference code, while the model weights (v2.5) are separately downloaded from Hugging Face Hub on first use. To ensure reproducibility, the package version is pinned to 6.3.2 (pip install tabpfn==6.3.2) in the experimental notebook.

The data was split into training and test sets using an 80:20 stratified random split, preserving class proportions across both partitions. Each model was trained on the same training set and evaluated on the same held-out test set. No cross-validation was performed for the final evaluation; the single train/test split was used for all reported metrics to ensure a direct, identical comparison across models. Figure 1 illustrates the overall research workflow, from data collection and target construction through the two parallel evaluation pipelines (baseline models with fixed hyperparameters versus zero-shot TabPFN on both raw and preprocessed input) to the unified evaluation stage.

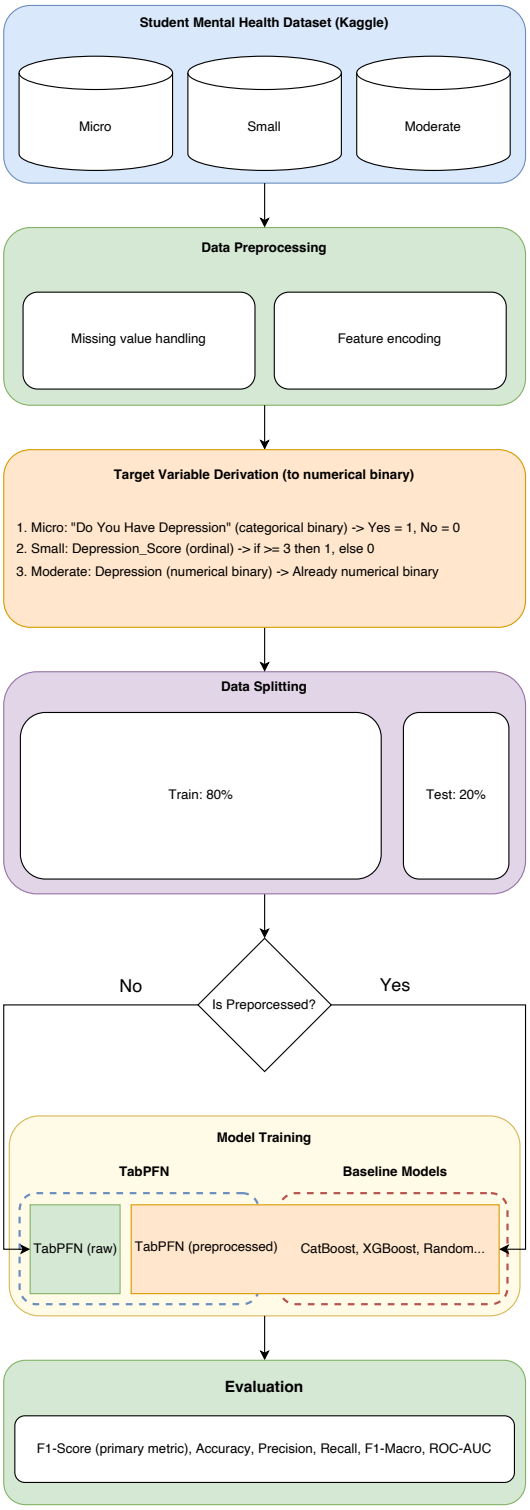

Figure 1. Research workflow diagram.

3.6. Ablation Study: Raw vs. Preprocessed Input

To evaluate TabPFN's claim of being able to handle raw, unprocessed tabular data, we conducted an ablation study. For each dataset, TabPFN was evaluated twice: once on the preprocessed data (with label encoding and standard scaling applied, matching the input given to the baseline models), and once on the raw data as downloaded from Kaggle, with no preprocessing applied beyond the train/test split and target variable construction. This ablation directly tests the practical ease-of-use advantage of TabPFN: if raw and preprocessed performance are comparable, researchers can skip the preprocessing pipeline entirely.

3.7. On the Absence of Statistical Testing

We note that this study does not employ formal statistical significance tests (e.g., McNemar's test or paired bootstrap) to compare model performance. This decision is deliberate and grounded in two considerations. First, for Dataset 1 (N_test = 21), the test set is too small for most statistical tests to have adequate power; the difference between the top-performing models corresponds to only one or two additional correct predictions, and any statistical test on such a sample would yield unreliable p-values. Second, for Datasets 2 and 3, we use a single train/test split rather than repeated cross-validation, which means there is no distributional basis for computing test statistics. The results should therefore be interpreted as point estimates of performance under a single experimental condition, not as generalisable claims of statistical superiority. Future work with larger test sets or repeated random splits would enable formal statistical comparison.

4. Results and Analysis

This section presents the comparative results across the three datasets, followed by the ablation study results.

4.1. Performance Comparison

Dataset 1: Micro-Sample (IIUM, N = 101). On the smallest and most imbalanced dataset, TabPFN achieved the highest performance across nearly all metrics. As shown in Table 3, TabPFN attained an F1-Score of 0.727 and Accuracy of 0.857, outperforming the best baselines, CatBoost and Random Forest, which both achieved an F1-Score of 0.667 and Accuracy of 0.810. SVM achieved an F1-Score of only 0.600 despite an Accuracy of 0.810, reflecting its tendency to favour the majority class (Precision = 1.000 but Recall = 0.429). Naive Bayes performed poorly overall (F1 = 0.462), though it was the only model to achieve high Recall (0.857), producing many false positives in the process. Figure 2 visualises the F1-Score comparison across all models on Dataset 1, clearly illustrating TabPFN's advantage in the micro-sample regime.

Table 3. Final performance on dataset 1: IIUM Student Mental Health (Micro-Sample). Acc: Accuracy; Prec: Precision; Rec: Recall; F1: F1-Score; F1-M: F1-Macro; AUC: ROC-AUC.

Model	F1	Acc	Prec	Rec	F1-M	AUC
TabPFN	**0.727**	**0.857**	**1.000**	0.571	**0.815**	**0.847**
CatBoost	0.667	0.810	0.800	0.571	0.767	0.776
Random Forest	0.667	0.810	0.800	0.571	0.767	0.776
XGBoost	0.615	0.762	0.667	0.571	0.721	0.765
SVM	0.600	0.810	**1.000**	0.429	0.738	0.776
Naive Bayes	0.462	0.333	0.316	**0.857**	0.293	0.464

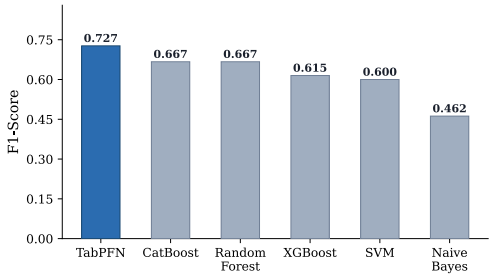

Figure 2. F1-Score comparison for Dataset 1.

Dataset 2: Small-Sample (Assessments, N = 7,022). On the small-sample dataset, the performance advantage shifts. As shown in Table 4, all models performed comparably, with F1-Scores tightly clustered. XGBoost achieved the highest F1-Score (0.432), followed by CatBoost (0.423). TabPFN achieved an F1-Score of 0.393, comparable to SVM (0.400) and Random Forest (0.393). The highest Accuracy was achieved by TabPFN (0.628), marginally above CatBoost (0.616). All models showed relatively low Recall on this dataset, suggesting that the binarised depression target was difficult to predict for all approaches. Figure 3 presents the corresponding F1-Score comparison for Dataset 2, showing the tightly clustered performance across models.

Table 4. Final performance on dataset 2: Mental Health Assessments (Small-Sample). Acc: Accuracy; Prec: Precision; Rec: Recall; F1: F1-Score; F1-M: F1-Macro; AUC: ROC-AUC.

Model	F1	Acc	Prec	Rec	F1-M	AUC
XGBoost	**0.432**	0.582	0.544	**0.358**	0.551	0.566
CatBoost	0.423	0.616	0.633	0.318	**0.567**	0.579
SVM	0.400	0.617	0.656	0.287	0.559	0.564
Naive Bayes	0.398	0.626	0.693	0.279	0.563	0.575
TabPFN	0.393	**0.628**	**0.713**	0.271	0.563	0.565
Random Forest	0.393	0.617	0.662	0.279	0.557	**0.586**

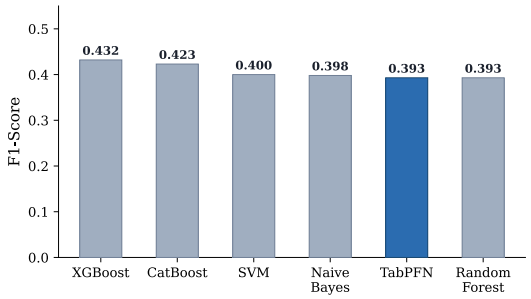

Figure 3. F1-Score comparison for Dataset 2.

Dataset 3: Moderate-Sample (Depression, N = 27,901). The moderate-sample dataset confirms the trend. As shown in Table 5, CatBoost achieved the highest F1-Score of 0.869, followed by XGBoost (0.865), SVM (0.864), TabPFN (0.864), and Random Forest (0.862). The top five models are clustered within a narrow 0.007 F1-Score range, indicating that once sufficient data is available, the choice of model has minimal impact on performance. TabPFN remains fully competitive (its F1-Score of 0.864 is within 0.005 of the best model), but it does not hold a clear advantage. Naive Bayes failed substantially on this dataset (F1 = 0.002), likely due to violated independence assumptions in the feature set. Figure 4 displays the F1-Score comparison for Dataset 3, confirming the convergence of all models except Naive Bayes within a narrow performance band.

4.2. Ablation Study: Raw vs. Preprocessed Input for TabPFN

Table 6 (next page) presents the results of the ablation study comparing TabPFN's performance on raw and preprocessed input.

The results are notable: across all three datasets, TabPFN produced identical or near-identical F1-Scores regardless of whether the input was preprocessed or raw. On Datasets 1 and 2, the F1-Scores were exactly identical between the two conditions. On Dataset 3, the difference was negligible (0.864 vs. 0.864). Minor differences were observed only in ROC-AUC on Dataset 1 (0.847 vs. 0.837), which is within the margin expected from minor numerical differences in probability outputs. This finding has a direct practical implication: for researchers working with micro-sample student mental health survey data, TabPFN can be applied directly to raw data without any preprocessing pipeline (no label encoding, no scaling, no imputation) and achieve comparable results. This substantially lowers the technical barrier to entry for non-ML-specialist researchers.

Table 5. Final performance on dataset 3: Student Depression (Moderate-Sample). Acc: Accuracy; Prec: Precision; Rec: Recall; F1: F1-Score; F1-M: F1-Macro; AUC: ROC-AUC.

Model	F1	Acc	Prec	Rec	F1-M	AUC
CatBoost	**0.869**	**0.844**	0.851	0.888	**0.837**	**0.913**
XGBoost	0.865	0.840	**0.853**	0.878	0.834	0.909
SVM	0.864	0.837	0.844	0.885	0.830	0.908
TabPFN	0.864	0.838	0.850	0.879	0.832	0.911
Random Forest	0.862	0.832	0.830	**0.897**	0.824	0.910
Naive Bayes	0.002	0.415	0.800	0.001	0.294	0.819

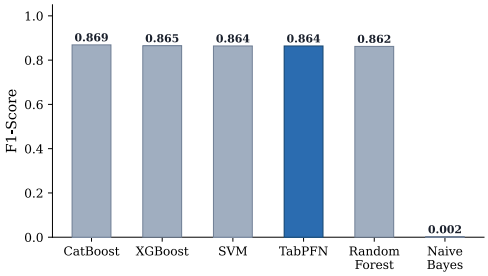

Figure 4. F1-Score comparison for Dataset 3.

5. Discussion and Conclusion

This section discusses the key findings of our comparative study, examines the practical implications for mental health researchers, acknowledges the limitations of the current work, and proposes directions for future research.

5.1. Key Findings

The empirical results of this study present a clear and practical narrative regarding model selection for supervised tabular student depression prediction data. Our central hypothesis was supported: the Tabular Prior-data Fitted Network demonstrates its clearest advantage in the micro-sample, imbalanced data regime.

On Dataset 1 (N = 101), TabPFN achieved the highest F1-Score of 0.727, a meaningful improvement over the best baselines (CatBoost and Random Forest, both at 0.667). This result, while based on a small test set (N_test = 21) and therefore subject to the caveats discussed in Section 3.7, is consistent with the design rationale of TabPFN: its pre-trained prior provides a strong inductive bias that is most beneficial when the data itself provides insufficient signal for training a model from scratch [7], [14].

On the small-sample Dataset 2 (N = 7,022) and moderate-sample Dataset 3 (N = 27,901), the performance advantage shifted to the tuned baseline classifiers, though the differences were small. On Dataset 2, XGBoost led with an F1-Score of 0.432, while TabPFN achieved 0.393. On

Table 6. TabPFN ablation: raw vs. preprocessed input

Dataset	Input	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Train Time (s)
Dataset 1	Preprocessed	0.857	1.000	0.571	0.727	0.847	5.66
Dataset 1	Raw	0.857	1.000	0.571	0.727	0.837	1.25
Dataset 2	Preprocessed	0.628	0.713	0.271	0.393	0.565	9.78
Dataset 2	Raw	0.628	0.713	0.271	0.393	0.563	9.96
Dataset 3	Preprocessed	0.838	0.850	0.879	0.864	0.911	33.05
Dataset 3	Raw	0.838	0.850	0.879	0.864	0.911	32.34

Dataset 3, CatBoost achieved the highest F1-Score of 0.869, with TabPFN at 0.864, a difference of only 0.005. These findings are consistent with the established literature on TabPFN's design niche [7], [14] and confirm that TabPFN is not intended to replace tuned ensembles on larger, data-rich tasks, but rather to provide a strong, zero-shot alternative when data is limited.

The ablation study further revealed that TabPFN's performance was robust to the absence of preprocessing. Across all three datasets, raw and preprocessed inputs yielded identical or near-identical F1-Scores, confirming TabPFN's practical ease of use for researchers who may lack the ML expertise to design appropriate preprocessing pipelines [7], [14].

5.2. Practical Implications

These findings yield a straightforward decision framework for mental health researchers:

When the available dataset is micro-sample (on the order of 100 instances), TabPFN should be the first model considered. It requires no hyperparameter tuning, no feature preprocessing, and can be run on freely accessible hardware (Google Colab with a T4 GPU). Its zero-shot

nature means that results can be obtained in minutes, not hours, making it particularly suitable for pilot studies, preliminary analyses, or settings where rapid iteration is needed.

When the dataset is of small to moderate size (thousands of instances or more), traditional ensembles such as CatBoost, XGBoost, or Random Forest are likely to be competitive or marginally superior. In these regimes, the investment in hyperparameter tuning is more reliable and worthwhile, as the larger validation sets provide more stable performance estimates. Even in these cases, however, TabPFN can serve as a useful calibration baseline: if a tuned model fails to outperform zero-shot TabPFN, this may indicate issues with the tuning process or the feature set.

5.3. Limitations

Several limitations should be acknowledged. First, the depression indicators in these datasets represent self-reported measures or ordinal scale cut-offs rather than clinical diagnoses derived from standardised instruments (e.g., PHQ-9, BDI-II). While such measures have demonstrated utility for population-level screening [19], the predictions made by any model trained on these features should not be interpreted as clinical diagnostic classifications.

Second, the study uses a single train/test split per dataset rather than repeated cross-validation or bootstrapped confidence intervals. This means the reported metrics are point estimates, and the performance differences, particularly on Dataset 1 with only 21 test instances, may not generalise across different splits. Future work should employ repeated random sub-sampling or k-fold cross-validation to assess variance.

Third, no formal statistical significance tests were performed to compare model performances (as discussed in Section 3.7). The reported improvements should be interpreted as descriptive, not as statistically confirmed effects.

Fourth, interpretability was not assessed. While TabPFN provides predictions, it does not natively produce feature importance scores or model explanations. For clinical or policy applications, coupling TabPFN with post-hoc interpretability methods (e.g., SHAP) would be an important next step.

5.4. Future Directions

Future research should pursue several directions. First, integrating model-agnostic interpretability methods such as SHAP would address the clinical requirement for explainable predictions. Second, prospective validation on institutional data, rather than retrospective Kaggle datasets, would provide stronger evidence for deployment. Third, expanding the comparison to include other emerging tabular models (e.g., TabICL, LimiX) would position the benchmark within the broader landscape of tabular foundation models.

Fourth, strengthening the statistical foundation of the comparison is essential. The present study relies on a single stratified train/test split, which yields point estimates but provides no measure of variance or basis for formal hypothesis testing. Future work should employ repeated stratified k-fold cross-validation (e.g., 10-fold repeated 10 times) to obtain distributional performance estimates, enabling the computation of confidence intervals and the application of appropriate

statistical tests such as the corrected paired t-test or the Wilcoxon signed-rank test. This is particularly important for Dataset 1, where the small test set (N_test = 21) means that a single additional correct or incorrect prediction can shift F1-Score by several percentage points. Repeated evaluation over multiple folds would clarify whether TabPFN's observed advantage on micro-sample data is robust across different data partitions or an artefact of the specific split used.

Fifth, future studies should evaluate TabPFN on datasets derived from clinically validated instruments such as the PHQ-9 [19] or the Beck Depression Inventory (BDI-II), rather than the self-reported binary indicators and ordinal scales used in the present work. Clinical instruments offer standardised severity thresholds grounded in diagnostic criteria, which would enable more meaningful target variable construction and allow direct comparison with established screening benchmarks. Such alignment would be a necessary step toward translating the predictive performance demonstrated here into tools that are actionable within university counselling and clinical settings.

References

[1] World Health Organization, "Depressive disorder (depression)," WHO Fact Sheet. Accessed: May 20, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[2] C. El Morr *et al.*, "Predictive machine learning models for assessing Lebanese university students' depression, anxiety, and stress during COVID-19," *J. Prim. Care Community Health*, vol. 15, p. 21501319241235588, 2024, doi: 10.1177/21501319241235588.

[3] M. I. H. Nayan, M. S. G. Uddin, M. I. Hossain, M. M. Alam, M. A. Zinnia, and I. Haq, "Comparison of the performance of machine learning-based algorithms for predicting depression and anxiety among university students in Bangladesh: A result of the first wave of the COVID-19 pandemic," *Asian J. Soc. Health Behav.*, vol. 5, no. 2, pp. 75–84, 2022, doi: 10.4103/shb.shb_38_22.

[4] W. Li, Z. Zhao, D. Chen, Y. Peng, and Z. Lu, "Prevalence and associated factors of depression and anxiety symptoms among college students: A systematic review and meta-analysis," *J. Child Psychol. Psychiatry*, vol. 63, no. 11, pp. 1222–1230, Nov. 2022, doi: 10.1111/jcpp.13606.

[5] R. Ahuja and A. Banga, "Mental stress detection in university students using machine learning algorithms," *Procedia Comput. Sci.*, vol. 152, pp. 349–353, 2019, doi: 10.1016/j.procs.2019.05.007.

[6] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[7] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," presented at the Proceedings of the 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, May 2023. [Online]. Available: https://openreview.net/forum?id=cp5PvcI6w8_.

[8] N. Hollmann *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, Jan. 2025, doi: 10.1038/s41586-024-08328-6.

[9] M. S. Islam, "Student mental health." Kaggle, 2020. Accessed: Nov. 4, 2025. [Online]. Available: https://www.kaggle.com/datasets/shariful07/student-mental-health.

[10] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[11] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1st ed. New York, NY, USA: Springer-Verlag, 1995. doi: 10.1007/978-1-4757-2440-0.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[13] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," presented at the Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), Geneva, Switzerland, Sep. 2013, pp. 671–676. doi: 10.1109/ACII.2013.117.

[14] L. Grinsztajn *et al.*, "TabPFN-2.5: Advancing the state of the art in tabular foundation models," Nov. 2025, *arXiv*. doi: 10.48550/arXiv.2511.08667.

[15] N. Erickson *et al.*, "TabArena: A living benchmark for machine learning on tabular data," presented at the Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, San Diego, CA, USA, Dec. 2025. [Online]. Available: https://openreview.net/forum?id=jZqCqpCLdU.

[16] D. Holzmüller, L. Grinsztajn, and I. Steinwart, "Better by default: Strong pre-tuned MLPs and boosted trees on tabular data," presented at the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, Dec. 2024, pp. 26577–26658. doi: 10.52202/079017-0837.

[17] S. Müller, N. Hollmann, S. Pineda Arango, J. Grabocka, and F. Hutter, "Transformers can do Bayesian inference," presented at the Proceedings of the 10th International Conference on Learning Representations (ICLR), Apr. 2022. [Online]. Available: https://openreview.net/forum?id=KSugKcbNf9.

[18] P. Wang *et al.*, "Harnessing small-data machine learning for transformative mental health forecasting: Towards precision psychiatry with personalised digital phenotyping," *Med Res.*, vol. 1, no. 2, pp. 226–238, 2025, doi: 10.1002/mdr2.70017.

[19] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," presented at the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, QC, Canada, Dec. 2018, pp. 6639–6649.

[21] sonia22222, "Students mental health assessments." Kaggle. Accessed: Nov. 4, 2025. [Online]. Available: https://www.kaggle.com/datasets/sonia22222/students-mental-health-assessments.

[22] S. Opeyemi, "Student depression dataset." Kaggle, Nov. 2024. Accessed: Nov. 4, 2025. [Online]. Available: https://www.kaggle.com/datasets/hopesb/student-depression-dataset.