Artificial Intelligence Applications And Innovations Vol. 2 No. 2 (2025)

Article history:
Received 28 June 2026
Revised 3 September 2026
Accepted 7 September2026
Published online 1 April 2025

STEM-Net: An Explainable Multi-Task Hybrid Architecture Combining Tabular Transformers, Gradient-Boosted Trees and Ensemble Learning for the Simultaneous Prediction of Eight Psychological and Academic Problems in High-School Students

AbdolRahim. Papi[1], Maseud. Rahgozar[2*], Habibollah. Arasteh Rad[3], Arshia. Badi[3], MohammadHossein. Rezvani[4]

1. Ph.D. Candidate in Information Technology at Aras International Campus – University of Tehran, Iran
2. School of Electrical and Computer Engineering – University of Tehran, Aras International Campus – University of Tehran, Iran
3. Institute for Advanced Applied Studies and Research – University of Tehran, Aras International Campus – University of Tehran, Iran
4. Department of Computer and Information Technology Engineering, Qa.C., Islamic Azad University, Qazvin, Aras International Campus – University of Tehran, Iran

*** Corresponding author email address: rahgozar@ut.ac.ir**

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Papi, A., Rahgozar, M., Arasteh Rad, H., Badi, A., Rezvani, M. (2025). STEM-Net: An Explainable Multi-Task Hybrid Architecture Combining Tabular Transformers, Gradient-Boosted Trees and Ensemble Learning for the Simultaneous Prediction of Eight Psychological and Academic Problems in High-School Students. *Artificial Intelligence Applications and Innovations, 2*(2), 27- 47.
https://doi.org/10.61838/jaiai.2.2.3

A B S T R A C T

Psychological and academic problems in students rarely occur in isolation: depression, anxiety, low self-esteem and addiction proneness tend to appear together, and modelling each of them separately wastes the signal they share. This study introduces the STEM-Net framework. Data were collected from 1,713 high-school students using eight validated questionnaires; after synthetic minority oversampling (SMOTE) the set grew to 5,139 records, and labels were coded at several risk levels based on the official cut-off points of each instrument. This multi-task hybrid architecture predicts all eight problems simultaneously, each at the risk levels defined by its own instrument. The base layer contains five heterogeneous learners in three branches: a multi-task tabular transformer whose attention body is shared across the eight targets, gradient-boosted tree models trained per target, and a pre-trained network. Before fine-tuning, the body of the deep branch is pre-trained on roughly eighty-eight thousand public respondents to two psychometric instruments, and the outputs of the five learners are combined through an automatic choice between stacking and simple averaging. To prevent item-to-score leakage, the items of the target scale are removed from the feature space of that task. On the test set STEM-Net reaches a macro-average accuracy of 0.9277, an F1 of 0.9034 and a ROC-AUC of 0.9844. A four-method explainability analysis shows that in every one of the eight tasks the most informative source is another one of the scales, and that addiction proneness and suicidal ideation are mutually predictive. Building on these features, a web-based screening system with a short questionnaire was designed that reports, for each student, the predicted risk level together with the features that most influenced it.

Keywords: *student mental health; multi-task learning; tabular transformer; ensemble learning; explainable artificial intelligence.*

1. Introduction

Adolescent mental health is a priority for education systems. The World Health Organization estimates that about one in seven adolescents aged 10 to 19 lives with a diagnosable mental disorder, and that a large share of these cases is never identified [1]. School is the most natural place for early screening, but a school screening tool has to be fast, cheap and explainable at the same time.

In recent years, educational data mining and learning analytics have produced many predictive models for students' academic and behavioural outcomes [2, 3]. Applying them to school mental health, however, runs into at least four serious methodological challenges, which the present study addresses directly.

A primary impediment in current mental health screening frameworks lies in the high-dimensional heterogeneity of psychometric assessments, which often overwhelms conventional machine learning estimators. Compounding this difficulty, existing architectures frequently struggle to balance representations across diverse, co-occurring psychological indicators without succumbing to feature suppression. Further complicating the modeling landscape is the severe class imbalance inherent to high-risk clinical tiers, where false negatives carry critical ethical liabilities. Finally, the opaque 'black-box' nature of high-capacity ensembles poses a persistent barrier to practical institutional deployment, demanding rigorous, multi-perspective model explainability before automated outputs can be entrusted to clinical practitioners.

The present study offers an integrated answer to these challenges in the form of the STEM-Net framework. Its main contributions can be summarised as follows:

- Adding a pre-training stage: the body of the deep network is first trained on public data from two psychometric questionnaires with nearly eighty-eight thousand respondents, then continues self-supervised training on all variables of the local data, and is finally fine-tuned on the study data with the body frozen and separate learning rates for the different layers.
- A base layer of five models in three branches, each with a different learning mechanism: a tabular transformer that learns all eight targets simultaneously, several gradient-boosted tree models each focused on a single target, and a pre-trained network.
- Combining model outputs separately for each target: for every task the framework decides automatically whether to use stacking or simple averaging, because stacking does not always give the better result.
- Labelling taken directly from the official manual of each questionnaire: instead of standardising the number of levels, the two- to five-level distinctions of each instrument are preserved, which turns the problem into an ordinal classification with twenty-nine classes in total. A threshold calibration step is also added after training.
- A four-method explainability module: tree Shapley values and permutation importance, which apply to all models, plus integrated gradients and attention-weight readout, which are specific to the deep network. All of these computations are performed out-of-fold so that the results are trustworthy.
- Deployment of the framework as a web-based screening system whose short questionnaire is built from the important features identified by the explainability module, and which gives every student, alongside the prediction, a personalised explanation of the factors behind that result.

Unlike our previous benchmark (SAFE-Net), which operated under compressed uniform class boundaries, STEM-Net preserves the full granularity of official psychometric manuals (ranging from 2 to 5 ordinal risk tiers across 8 domains). While direct performance comparisons between STEM-Net and SAFE-Net reflect both architectural upgrades and changes in target class definitions, our component-wise evaluation demonstrates that each methodological addition—specifically self-supervised tabular pre-training, multi-task cross-attention, and heterogeneous ensemble stacking—yields consistent and measurable gains in discriminative power.

The rest of the paper is organised as follows: Section 2 covers the background and the research gap; Section 3 the materials and methods; Section 4 the results; Section 5 the discussion and limitations; and Section 6 the conclusion.

2. Background

The first wave of research on predicting student mental health focused on classical classification algorithms and relied mostly on academic and behavioural outcomes [3]; one study, for example, showed that family and social variables play a considerable role in predicting depression

and anxiety [2]. This work was usually single-target and ignored the co-occurrence of psychological problems, even though such co-occurrence is the rule rather than the exception [4]. Multi-task learning is a direct response to this issue [5], and applying it to the simultaneous prediction of depression, anxiety and stress has given promising results [6].

For tabular data, the review by Borisov and colleagues showed that deep networks are usually weaker than gradient-boosted tree models unless their architecture is designed for tabular data [7]. A transformer that turns each feature into a separate token and learns the relationships through attention performs more competitively [8] and is also well suited to multi-task learning.

Another line of work is in-context learning on tabular data. TabPFN, pre-trained on millions of synthetic datasets, does not learn new parameters at all: it takes the training data as context and predicts directly, an approach designed for small datasets [9]. Combining it with tree models increases the diversity of the base layer.

In transfer learning, the common recipe is to train the model self-supervised on unlabelled data first and then fine-tune it. Public archives of raw questionnaire responses are a good source for this stage, because the way people answer Likert items is very similar across instruments. Since only the feature-to-token embedding of a tabular transformer is specific to a given feature, the knowledge stored in the body transfers to all subsequent targets.

In earlier work by the present authors, the multi-task framework SAFE-Net was introduced for the simultaneous prediction of these same eight constructs on this same dataset: a two-level stacked framework with four relatively homogeneous base learners, which formulated the per-target feature-masking mechanism that prevents item-to-score leakage — a mechanism retained in the present study. That work reached an accuracy of 0.8637 and an F1 of 0.7934, but had three limitations: the base learners all came from closely related families and were trained from scratch; the output of all eight instruments was compressed into three levels, losing clinically meaningful distinctions; and for academic motivation and family functioning the F1 dropped to 0.5969 and 0.6846. The present study keeps the same feature masking but answers these limitations with a different model structure, a different data source and a different labelling scheme, and uses that earlier framework as its comparison baseline. To prevent data leakage, the dataset was first partitioned into training (80%) and holdout test (20%) sets using stratified sampling across target risk categories. The SMOTE algorithm was applied strictly and exclusively to the training partition to balance class representation (expanding training instances from the training split while keeping natural distributions intact in the test split). The test set remained completely pristine and unaugmented throughout all experiments. All hyperparameter tuning and model selection routines were conducted via 5-fold stratified out-of-fold cross-validation solely on the training data.

Stacked generalisation, introduced by Wolpert, feeds the outputs of several base models into a higher-level model and usually outperforms even the best base model [10]. But "usually" does not mean "always": when there are few out-of-fold samples, the meta-learner overfits and simple averaging is more stable. In STEM-Net this consideration becomes an automatic per-target choice.

In explainability, SHAP provides a coherent framework for attributing the contribution of each feature through Shapley values [11], and permutation importance is a model-agnostic complement to it. For deep networks, integrated gradients measures each feature's contribution by integrating the gradient along the path from a baseline sample to the target sample [12], and reading out the attention weights shows which tokens the network attended to. Reporting all four methods together makes it possible to check how well they agree.

2.1. Research gap, objective and research questions

Taken together, this background reveals three overlapping gaps. First, in school psychometric data — where the sample is inherently small — almost no study uses the available public data for pre-training. Second, the base layer of most ensemble architectures is drawn from closely related methods and therefore creates no real diversity. Third, compressing clinical levels into a uniform three-level scheme destroys part of the discriminative information before training even starts.

The main objective of this study is to design and evaluate the STEM-Net framework: a framework that fills these three gaps while leaving intact the feature-masking mechanism that is the most important safeguard of label validity. The research questions are:

- Can transfer pre-training on public psychometric data, combined with an ensemble of five heterogeneous learners, achieve performance reliable enough for initial screening in all eight

tasks — and do so on a multi-level distinction that is harder than the common two- or three-level schemes?

- Is an automatic per-target choice between stacking and averaging a defensible option compared with imposing a single combination strategy on all tasks?
- To what extent can post-training threshold calibration shift the balance between precision and recall in imbalanced tasks?
- Which scale blocks contribute most to predicting each task, and does the pattern of cross-scale dependency remain stable when the architecture and the label levels change?

On this basis, two hypotheses are proposed.

First: transfer pre-training and greater diversity in the base layer improve performance across all targets, not only those whose instrument is present in the public data.

Second: if the dependency pattern between the different parts of the questionnaire reflects genuine co-occurrence of psychological problems, it should remain stable when the model structure and the labelling scheme change.

3. Materials and Methods

In terms of purpose this is applied research, and in terms of data collection it is a descriptive survey study with a quantitative approach. The following sections describe, in order, the population and sample, the architecture of the framework, the measurement instruments and labelling logic, the data processing pipeline, feature masking, transfer pre-training, training and combination, calibration, and the evaluation metrics.

3.1. Population, sample and data structure

Data were collected through a web-based questionnaire system. The link was distributed in high schools in Tehran — public and private, girls' and boys' schools — and completion was voluntary and self-reported, so the sampling is convenience sampling. This is the same dataset used in the authors' previous study, with no new records added, so any difference in the results comes purely from the change in method.

After cleaning, the final dataset had 243 columns. The score and interpretation columns of the eight scales were used as the basis for labelling and excluded from the feature space. The set of 1,713 records grew to 5,139 records after data augmentation. The final feature space has 217 columns: 192 raw items from the eight questionnaires and 25 contextual and demographic variables. Table 1 reports the composition of this space by block.

Table 1. Composition of the feature space by scale block

Block	Instrument	No. of features	Features permitted for the corresponding target
Self-esteem	RSES	10	207
Depression	BDI-II	15	202
Anxiety	DASS	7	210
Stress	DASS	7	210
Family functioning	FAD	60	157
Addiction proneness	IAPS	41	176
Academic motivation	HEMS	33	184
Suicidal ideation	BSSI	19	198
Contextual and demographic	—	25	—
Total	—	217	—

Source: extracted by the authors from the metadata of the trained model.

The last column of Table 1 is a direct consequence of the masking mechanism: when a given target is being predicted, the items of that target's own block are dropped from the feature space. Family functioning, the instrument with the most items in the study, therefore has the narrowest subspace, with 157 features.

Note also that the block of behavioural response-pattern features has been removed from the feature space. This decision was based on a finding from the authors' previous study on the same data: the contribution of this block never exceeded 0.0408 and was even negative in two tasks. Removing it made the feature space lighter without losing any signal.

3.2. . Architecture of the proposed framework

The STEM-Net architecture has four conceptual layers. Layer zero builds the permitted feature space for each target.

Layer one consists of five heterogeneous base learners in three branches with different inductive logic. Layer two passes their out-of-fold probabilities to the combination unit, which decides for each target between stacking and simple averaging. Layer three calibrates the output and hands it to the explainability module. Figure 1 shows this structure.

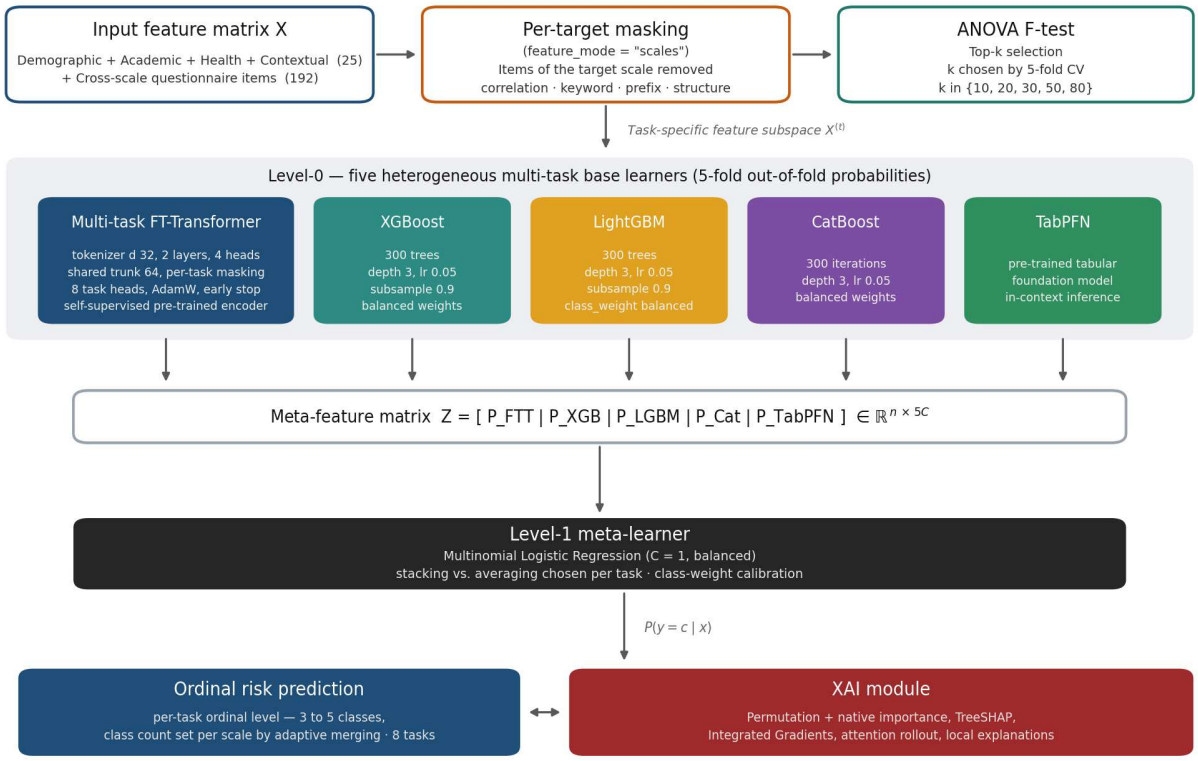

Figure 1. Overview of the proposed framework's architecture

Branch one: a multi-task transformer designed for tabular data. Each feature is turned into a numeric vector by its own embedding, a class token is appended to the sequence, and two attention blocks are applied to it; the output of the class token passes through a shared trunk and reaches eight classification heads. The key point is that the attention blocks and the shared trunk are identical for all eight targets, and it is precisely this sharing that makes knowledge transfer between tasks possible.

Branch two: a group of gradient-boosted tree models. XGBoost, LightGBM and CatBoost are each trained for every target on that target's permitted features. All three belong to the gradient boosting family, but they differ in how they grow trees, how they handle missing values and how they regularise, and these differences make their errors partly different from one another [13].

Branch three: a pre-trained network. This model was previously trained on millions of synthetic datasets and does no new learning at inference time; instead it receives the training data as context and makes its prediction. It is included in the base layer because it was designed for exactly the small-sample setting of this study.

3.3. *Measurement instruments and labelling logic*

In this field, the validity of any model depends above all on the validity of its labels. The eight target constructs were measured with eight self-report instruments whose Persian versions and official scoring manuals were obtained from the Azmoonyar Pouya system. Table 2 reports the number of items, response scale, score range and direction of each instrument; direct means that a higher score indicates greater risk, and reverse means that a higher score indicates a more favourable condition.

Table 2. Technical specifications of the eight measurement instruments

Instrument	Items	Response scale	Score range	Direction	Reference
Rosenberg Self-Esteem Scale (RSES)	10	Four-point, 0–3, with five reverse-scored items	0 to 30	Reverse	[14]
DASS-21 anxiety subscale	7	Four-point, 0–3; the sum is multiplied by two	0 to 42	Direct	[15]
DASS-21 stress subscale	7	Four-point, 0–3; the sum is multiplied by two	0 to 42	Direct	[15]
Beck Depression Inventory, second edition (BDI-II)	21	Four options, 0–3	0 to 63	Direct	[16]
McMaster Family Assessment Device (FAD)	60	Four-point, 1–4, with reverse scoring on healthy items	Mean of 1 to 4	Direct	[17]
Beck Scale for Suicide Ideation (BSSI)	19	Three-point, 0–2	0 to 38	Direct	[18]
Addiction Potential Scale (IAPS-41)	41 items: 36 scored and 5 lie-detection items	Four-point, 0–3	0 to 108	Direct	[19]
Academic Motivation Questionnaire	33	Five-point Likert, 1–5, with eleven reverse-scored items	33 to 165	Reverse	Harter (HEMS)

Source: official scoring manuals of the instruments, obtained from the Azmoonyar Pouya system.

The Rosenberg Self-Esteem Scale is a ten-item, single-factor instrument in which five items are reverse-scored [14]. For depression, anxiety and stress, the short twenty-one-item version was used; following the official manual, the sum of each subscale is multiplied by two so that it can be compared with the normative table of the long version. The depression subscale of this instrument was not used as a target, because depression was measured with the more specific Beck inventory [16].

The McMaster Family Assessment Device measures family functioning across seven dimensions; the score of each subscale is the mean of its items on a scale of one to four, and a higher score indicates less healthy functioning. Labelling was based on the general functioning score. The Beck Scale for Suicide Ideation measures the intensity of attitudes, behaviours and planning for an attempt over the past week, and its first five items serve a screening function [18]. The Iranian forty-one-item version of the Addiction Potential Scale measures active and passive proneness [19], and the thirty-three-item Academic Motivation Questionnaire is reverse-directed.

The usual practice in this field is to convert the continuous score of each scale into two or three levels using one or two cut-offs. In the present framework the label is read directly from the official interpretation column of the instrument and converted into an ordinal label, so that every level defined by the official manual is preserved in the label as well. For scales with a direct direction, the ordinal label is obtained from Equation (1):

$$y_t = \Sigma(m=1..M)\ I(s \geq c_m) \quad (1)$$

and for scales with a reverse direction, such as self-esteem and academic motivation, in which a lower score signals higher risk, Equation (2) applies:

$$y_t = \Sigma(m=1..M)\ I(s < c_{(M-m+1)}) \quad (2)$$

In these equations s is the raw score of the scale, c are the official cut-offs of the instrument, M is the number of cut-offs and I is the indicator function. Table 3 reports the levels of all eight tasks and their corresponding cut-offs.

Table 3. Label levels and official cut-offs for the eight tasks

Task	Instrument	Levels	Cut-offs	Level names
Self-esteem	RSES	3	20 and 10.5 (reverse direction)	High; Moderate; Low
Anxiety	DASS×2	5	5; 11; 18; 26.5	Normal; Mild; Moderate; Severe; Extremely severe
Stress	DASS×2	5	12; 18.5; 25.5; 33	Normal; Mild; Moderate; Severe; Extremely severe
Family functioning	FAD	2	2.4482	Good or moderate; Poor
Depression	BDI-II	4	13; 26.5; 38.5	Minimal; Mild; Moderate; Severe
Addiction proneness	IAPS	4	30.5; 60.5; 92.5	Low level; Moderate level; High level; Very high proneness
Academic motivation	HEMS	3	98.5 and 71 (reverse direction)	Very good; Moderate; Poor
Suicidal ideation	BSSI	3	7; 22	Low risk; High risk; Very high risk

Source: matched by the authors between the official instrument manuals and the rules encoded in the model metadata.

This choice makes the problem noticeably harder: compressing everything to three levels would give a 24-class problem, whereas preserving the official levels brings it to 29 classes, and for anxiety and stress the model has to distinguish among five severity bands. In exchange, the model's output speaks exactly the language that the official manual and the school counsellor already use, and no discriminative information is discarded before training.

3.4. Data processing pipeline

Figure 2 shows the complete data processing pipeline, from the raw output of the questionnaire system through to evaluation and explanation.

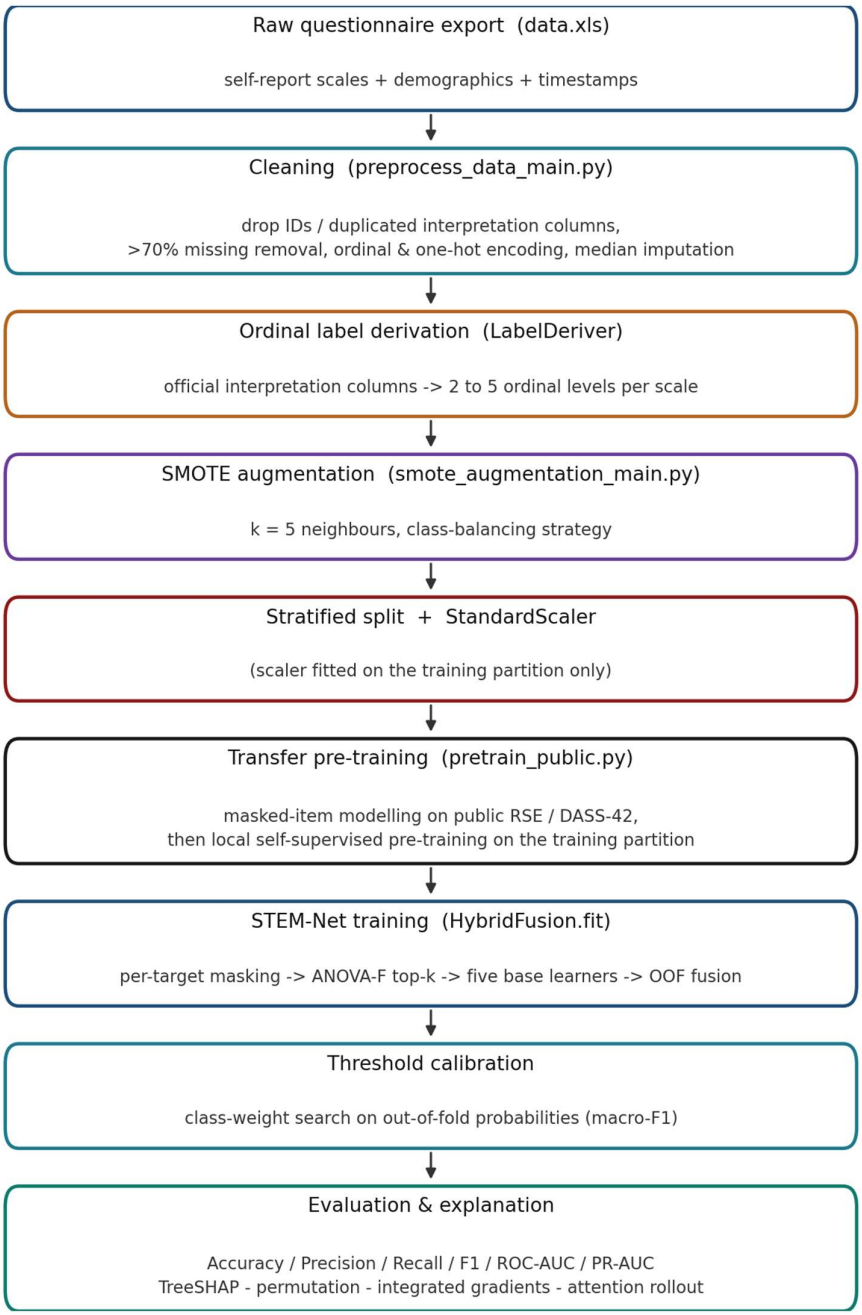

Figure 2. Pipeline for data processing, labelling, training and evaluation
Source: drawn by the authors.

3.4.1. Data cleaning and encoding

The data were cleaned first: identifier columns and columns that were low-information duplicates of the main scores were removed so that no redundant or collinear information remained, and any column with more than seventy per cent missing data was also dropped. Ordinal variables such as gender, economic status, parental education and Likert responses were converted to numbers, and nominal variables were one-hot encoded.

Remaining missing values were imputed with the median of the same column. The median of every feature is stored in the model metadata so that exactly the same value is used at inference time and the model behaves identically in training and in deployment.

3.4.2. Class balancing and data augmentation

To reduce the effect of class imbalance, synthetic minority oversampling (SMOTE) was used: for each minority sample, one of its k nearest neighbours in the same class is selected and a synthetic sample is generated by linear interpolation according to Equation (3) [20]:

$$x_{new} = x_i + \lambda\,(x_{nn} - x_i), \quad \lambda \sim U(0, 1) \tag{3}$$

In this implementation the number of neighbours was set to five, the augmentation factor to three, and the sampling strategy to full class balancing.

3.4.3. Data splitting and scaling

The data were split into training and test sets in an eighty-twenty ratio, stratified so that the proportion of each class is preserved in both parts. Standard scaling is fitted on the training data only and applied to the test data with the same parameters; the same rule is followed for feature selection and threshold tuning, meaning that no stage of training has access to the test data.

3.5. Per-target feature masking and feature selection

One of the methodological pillars of this study is the feature-masking mechanism. Three options were possible: using all items, which leads to direct item-to-score leakage; using only contextual variables, which carry no psychometric information; and the middle option, which was chosen, in which only the items of the target's own scale are removed for that target. The model therefore cannot reconstruct the target score from its constituent items, but still benefits from the correlations between constructs; this same mechanism is what carries knowledge from one task to another. The number of permitted features per target is: 210 for anxiety and stress, 207 for self-esteem, 202 for depression, 198 for suicidal ideation, 184 for academic motivation, 176 for addiction proneness and 157 for family functioning.

After masking, univariate feature selection is performed using the F statistic of a one-way analysis of variance, which expresses the ratio of between-group variance to within-group variance; this step and most of the classical components of the framework are implemented with the scikit-learn library [21].

$$F_j = MS_{between} / MS_{within} \tag{4}$$

The number of retained features, k, is not fixed: for each target, and only during training, it is chosen from {10, 20, 30, 50, 80} by five-fold stratified cross-validation maximising macro F1, and for self-esteem and stress the grid was extended to {10, 20, 40, 60, 90, 130}. To keep the cost of this search down, a lightweight histogram gradient boosting model is used. The final values were 130 for self-esteem and stress and 80 for the other six tasks.

3.6. Transfer pre-training

This part is carried out in three stages.

The first stage is pre-training on public data. Two anonymous sets from a public archive of questionnaire responses were used: the Rosenberg Self-Esteem Scale with 47,974 respondents and the forty-two-item Depression Anxiety Stress Scales with 39,775 respondents; twenty-four items from these two were aligned with their Persian equivalents. Training was performed with two simultaneous objectives: reconstructing masked items (thirty-five per cent of the items) and predicting the severity level of each subscale from the items of the other subscales. The total loss is the weighted sum of these two objectives:

$$L_{pre} = L_{mask} + \alpha \cdot L_{severity} \tag{5}$$

The second stage is self-supervised training on the local data: the same masked-item reconstruction objective, this time over all 217 features and only on the rows of the training split. The importance of this stage is that items with no comparable public data — such as the items of the Beck Depression Inventory, the Family Assessment Device, addiction proneness, motivation and suicidal ideation — are also given good representations. The train/test boundary is respected at this stage.

The third stage is supervised fine-tuning on the study data. Two measures guard against catastrophic forgetting: first, the body of the network is frozen for the first eight epochs so that the classification heads, which start from random weights, can stabilise; second, once the body is unfrozen, its learning rate is set to one tenth of that of the heads:

$$\eta_{backbone} = 0.1 \cdot \eta_{heads} \quad (6)$$

3.7. *Branch training and the combination unit*

To prevent information leaking into the combination layer, the higher-level features are built from out-of-fold probabilities: the training data are split into five stratified folds and the prediction for each sample is produced by a model that has not seen that sample. For sample i and branch m:

$$Z^{(m)}[i, :] = \hat{h}^{(m, -k(i))}(x_i) \quad (7)$$

The input matrix of the combination unit is formed by horizontally concatenating the probability blocks of each branch. With C classes and five probability sources — one tabular transformer, three tree models and one pre-trained network — its dimension is five times the number of classes of that target:

$$Z = [\, Z^{(FTT)} \mid Z^{(xgb)} \mid Z^{(lgbm)} \mid Z^{(cat)} \mid Z^{(tabpfn)} \,] \in \mathbb{R}^{(n \times 5C)} \quad (8)$$

The usual practice in stacked generalisation is for the meta-model always to be a multinomial logistic regression. In the present framework, one of two options is chosen for each target. The first is stacking itself, which models the class probability distribution according to Equation (9):

$$P(y = c \mid x) = \exp(w_c^T z + b_c) / \Sigma_{j=1..C} \exp(w_j^T z + b_j) \quad (9)$$

The second option is a simple average of the probabilities of all branches. To choose between the two, the performance of stacking is measured by three-fold cross-validation on the same out-of-fold matrix and compared with the macro F1 of simple averaging, and whichever is better is selected. The logic is simple: when a class has few samples, the meta-model itself overfits and simple averaging is more stable.

To handle class imbalance, in models that support sample weighting each sample is weighted inversely to the number of samples in its class, because on such data a model tends to ignore the rare classes. This forces the model to reduce the error on rare classes as well and keeps the balance between precision and recall better across all classes:

$$w_i = n / (C \cdot n_{y(i)}) \quad (10)$$

The settings of the tree models are also tuned separately for each target: different combinations of depth and number of trees are evaluated by inner cross-validation and the best combination is selected. Table 4 shows the exact settings of all components, extracted directly from the configuration of the final model.

Table 4. Settings and hyperparameters of the framework's components

Component	Settings
Multi-task transformer	Latent dimension 32, two attention blocks, four attention heads, dropout 0.2, shared trunk with 64 neurons, eight separate classification heads
Deep branch training	AdamW optimiser, learning rate 0.001, weight decay 0.01, batch size 32, maximum 80 epochs, early stopping with patience of 12 epochs
Transfer pre-training	Mask ratio 0.35; body frozen for the first eight epochs; body learning rate 0.1× the head learning rate
XGBoost	300 trees, learning rate 0.05, max depth 3, row and column subsampling 0.9, histogram tree construction
LightGBM	300 trees, learning rate 0.05, max depth 3, row and column subsampling 0.9, balanced class weighting
CatBoost	300 iterations, learning rate 0.05, depth 3, automatic balanced class weighting
Tree hyperparameter tuning	Per-target selection among five configurations (depth, number of trees): (2, 200), (2, 400), (3, 150), (3, 300), (4, 200)
Pre-trained network	Maximum 500 input features, run on GPU where available
Combination unit	Automatic per-target choice between stacking (logistic regression, inverse regularisation coefficient 1, balanced weighting) and simple averaging; compared by three-fold cross-validation
Feature selection	ANOVA F statistic; k selected from {10, 20, 30, 50, 80}, and from {10, 20, 40, 60, 90, 130} for self-esteem and stress, with five-fold cross-validation
Out-of-fold validation	Stratified five-fold cross-validation with random shuffling
Threshold calibration	300 class-weight search iterations on out-of-fold probabilities, using macro F1
Explainability module	Five repetitions in permutation importance, top twenty features reported, 64 samples and 32 steps in integrated gradients
Random seed	42 for all components

Source: extracted by the authors from the saved configuration of the trained model.

3.8. *Threshold calibration*

The output of the combination unit is a probability vector, and the final decision is normally taken by a simple argmax, which in imbalanced tasks is systematically biased in favour of the majority classes. To correct this bias, a vector of class weights is searched for after training, so that the probability of each class is multiplied by its corresponding weight before the argmax:

(11) $\hat{y} = \text{argmax}_c \left(\gamma_c \cdot P(y = c \mid x) \right)$

The vector γ is tuned by 300 random search iterations, using only the out-of-fold probabilities of the training split, so as to maximise macro F1; the test data play no part at this stage, and results both before and after this tuning are reported.

3.9. *Evaluation metrics and explainability analysis*

All metrics are computed as macro averages over the classes of each task and on the test data. Overall accuracy is defined by Equation (12):

(12) $\text{Accuracy} = (1/n) \sum_{i=1..n} I(\hat{y}_i = y_i)$

Macro precision and macro recall are the averages of the per-class values:

(13) $\text{Precision}_{macro} = (1/C) \sum_c TP_c / (TP_c + FP_c)$, $\text{Recall}_{macro} = (1/C) \sum_c TP_c / (TP_c + FN_c)$

and macro F1 is computed as the average of the per-class F1:

(14) $F1_{macro} = (1/C) \sum_c 2 \cdot P_c \cdot R_c / (P_c + R_c)$

The area under the receiver operating characteristic curve and the area under the precision-recall curve were computed one-vs-rest for each class and then averaged.

The explainability module reports four methods at once. The permutation importance of feature j is obtained from the drop in performance after randomly shuffling that feature:

(15) $PI_j = (1/R) \sum_{r=1..R} [M(D) - M(D_{\pi(j)^{(r)}})]$

In the tree explanation method, each feature's contribution is computed from Shapley values and indicates how far each feature moves the prediction away from the base value [11]:

(16) $f(x) = \varphi_0 + \sum_{j=1..p} \varphi_j$

For the deep branch, integrated gradients computes the contribution of feature j by integrating the gradient of the output along the straight path between a baseline sample (x′) and the target sample (x) [12]:

(17) $IG_j(x) = (x_j - x'_j) \int_0^1 \partial F(x' + a(x - x')) / \partial x_j \, da$

The fourth method is reading out the attention weights to establish how much each input influenced the final result. Importantly, all of these computations are performed out-of-fold: each part of the data is explained by a model that has not seen it, so every row is explained exactly once and fairly.

Besides the effect of each feature on its own, the group effect of features is also computed: the contributions of features belonging to the same scale are summed so as to control for the correlation within that scale.

The reduction from the comprehensive 217-item battery to the abbreviated screening tool was executed through a three-stage filtering pipeline: (1) global feature importance ranking derived from the ensemble explainability consensus (combining TreeSHAP and Integrated Gradients), (2) removal of collinear and redundant items via pairwise correlation analysis ($r > 0.85$), and (3) clinical sanity checks ensuring that core diagnostic markers for all eight constructs remained represented. Importantly, while this abbreviated instrument exhibits high retrospective screening fidelity against the full assessment, it serves as an exploratory prioritized battery and has not yet undergone independent external psychometric standardization in a fresh cohort—a key objective for future longitudinal trials.

4. Results

4.1. *Experimental setup and assumptions*

Before reporting the results, the experimental setup and the assumptions governing the evaluation are collected in Table 5.

Table 5. Experimental setup and evaluation assumptions

Item	Value / setting
Type of research	Applied in purpose; descriptive survey with a quantitative approach in data collection
Population and sampling method	High-school students in Tehran (public and private girls' and boys' schools); voluntary convenience sampling
Data collection period	December 2025 to June 2026
Number of complete responses	1,713 responses
Number of columns in the final dataset	243 columns
Data augmentation	Synthetic minority oversampling with five neighbours and full class balancing; 5,139 records comprising 1,713 real and 3,426 synthetic records
Train/test split	Eighty-twenty, stratified
Target tasks	Eight tasks with two to five levels; 29 classes in total
Basis for labelling	The official interpretation column of each instrument, mapped to an ordinal label according to Equations (1) and (2)
Feature space	217 features, comprising 192 questionnaire items and 25 contextual variables
Per-target feature masking	Items of the target scale removed from the feature space, items of the other scales retained
Number of selected features per target	130 for self-esteem and stress; 80 for the other six tasks
Base learners	Five learners in three branches: a multi-task tabular transformer; a tree group of XGBoost, LightGBM and CatBoost; and the pre-trained TabPFN network
External knowledge source	47,974 respondents to the Rosenberg Self-Esteem Scale and 39,775 respondents to the Depression Anxiety Stress Scales
Meta-feature construction	Out-of-fold probabilities from stratified five-fold cross-validation
Combination unit	Automatic per-target choice between stacking and simple averaging
Evaluation metrics	Accuracy, macro precision, macro recall, macro F1, ROC-AUC and PR-AUC

Source: compiled by the authors from the Materials and Methods section and the saved configuration of the trained model.

4.2. Overall and per-task performance

Table 6 reports the performance of the proposed framework on the test set, broken down by the eight tasks. The macro-average accuracy was 0.9277, the mean F1 0.9034, the mean ROC-AUC 0.9844 and the mean PR-AUC 0.9588.

Table 6. Performance of the proposed framework on the test set, by task

Task	Levels	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Self-esteem	3	0.9410	0.8933	0.9136	0.9020	0.9783	0.9581
Anxiety	5	0.8646	0.8217	0.8388	0.8286	0.9757	0.9127
Stress	5	0.8865	0.8638	0.8792	0.8692	0.9801	0.9392
Family functioning	2	0.9367	0.9365	0.9370	0.9366	0.9811	0.9824
Depression	4	0.9192	0.9082	0.9114	0.9086	0.9820	0.9626
Addiction proneness	4	0.9454	0.9481	0.9514	0.9492	0.9926	0.9818
Academic motivation	3	0.9607	0.8917	0.8431	0.8646	0.9926	0.9436
Suicidal ideation	3	0.9672	0.9685	0.9681	0.9682	0.9932	0.9902
Macro average	—	0.9277	0.9040	0.9053	0.9034	0.9844	0.9588

Source: computed by the authors from the evaluation report of the trained model.

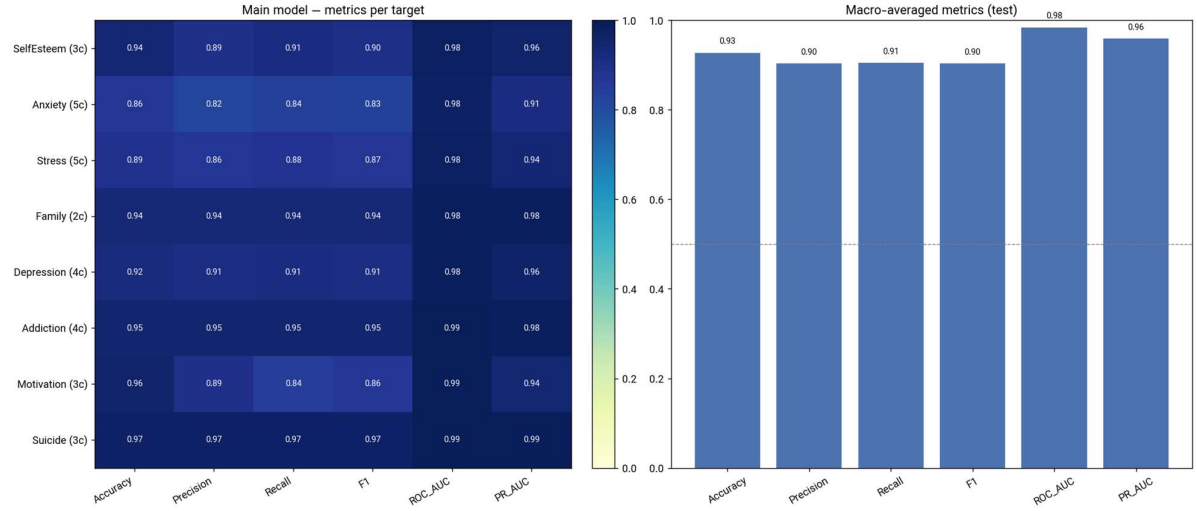

Figure 3. Comparison of six evaluation metrics across the eight target tasks and their macro averages

Source: output of the study's evaluation module.

Performance varies across the eight tasks, but not by much. The best result belongs to suicidal ideation with an F1 of 0.9682, followed by addiction proneness at 0.9492 and family functioning at 0.9366. The lowest is anxiety at 0.8286, a task with five severity levels whose adjacent boundaries are sometimes only two or three points apart. The gap between best and worst is only 0.1396, which means the framework is also balanced.

The ROC-AUC is above 0.975 in all eight tasks, meaning that the model is reliable at ranking probabilities even in tasks such as anxiety where its accuracy is lower, and that part of the error comes from the choice of threshold rather than from an inability to discriminate.

4.3. *Effect of threshold calibration*

Table 7 shows the four main classification metrics after class-weight tuning. The two area-under-curve metrics are not repeated in this table, because weight tuning changes only the decision rule and not the ordering of the probabilities.

Table 7. Performance of the proposed framework after threshold calibration

Task	Accuracy	Precision	Recall	F1	Change in F1
Self-esteem	0.9432	0.9594	0.8670	0.9070	+0.0050
Anxiety	0.8712	0.8513	0.8216	0.8332	+0.0046
Stress	0.8996	0.8792	0.8837	0.8809	+0.0117
Family functioning	0.9236	0.9251	0.9225	0.9233	−0.0133
Depression	0.9279	0.9228	0.9173	0.9192	+0.0106
Addiction proneness	0.9541	0.9568	0.9595	0.9580	+0.0088
Academic motivation	0.9629	0.8937	0.8444	0.8661	+0.0015
Suicidal ideation	0.9563	0.9595	0.9570	0.9582	−0.0100
Macro average	0.9299	0.9185	0.8966	0.9057	+0.0023

Source: computed by the authors from the evaluation report of the trained model.

Calibration has a positive but small effect: the mean F1 rises from 0.9034 to 0.9057 and the mean accuracy from 0.9277 to 0.9299. More important than the size is the direction of the change: mean precision rises from 0.9040 to 0.9185 while mean recall falls from 0.9053 to 0.8966. In other words, calibration makes the model more cautious: there are fewer false alarms, but some genuine cases are missed as well.

F1 improves in six of the eight tasks, the largest gain being 0.0117 for stress; but family functioning and suicidal ideation both decline. Recall for suicidal ideation falls from 0.9681 to 0.9570, and in a task where the cost of missing a high-risk individual far exceeds the cost of a false alarm, the uncalibrated results are the better choice.

4.4. *Class-level analysis*

Macro averages hide the differences within a task. Table 8 reports performance at the level of individual classes.

Table 8. Class-level performance of the proposed framework on the test set

Task	Level	Precision	Recall	F1
Self-esteem	High	0.9239	0.8673	0.8947
Self-esteem	Moderate	0.9560	0.9645	0.9602
Self-esteem	Low	0.8000	0.9091	0.8511
Anxiety	Normal	0.8817	0.8817	0.8817
Anxiety	Mild	0.5581	0.6857	0.6154
Anxiety	Moderate	0.8966	0.8667	0.8814
Anxiety	Severe	0.8202	0.8690	0.8439
Anxiety	Extremely severe	0.9521	0.8910	0.9205
Stress	Normal	0.9403	0.9000	0.9197
Stress	Mild	0.6610	0.8298	0.7358
Stress	Moderate	0.8228	0.8667	0.8442
Stress	Severe	0.9175	0.8641	0.8900
Stress	Extremely severe	0.9775	0.9355	0.9560
Family functioning	Good / moderate	0.9485	0.9286	0.9384
Family functioning	Poor	0.9244	0.9455	0.9348
Depression	Minimal	0.8974	0.9722	0.9333
Depression	Mild	0.8333	0.8784	0.8553
Depression	Moderate	0.9381	0.8426	0.8878
Depression	Severe	0.9639	0.9524	0.9581
Addiction proneness	Low level	0.9510	0.9444	0.9477
Addiction proneness	Moderate level	0.9231	0.8710	0.8963
Addiction proneness	High level	0.9182	0.9902	0.9528
Addiction proneness	Very high proneness	1.0000	1.0000	1.0000
Academic motivation	Very good	0.9610	0.9652	0.9631
Academic motivation	Moderate	0.9641	0.9641	0.9641
Academic motivation	Poor	0.7500	0.6000	0.6667
Suicidal ideation	Low risk	0.9653	0.9799	0.9726
Suicidal ideation	High risk	0.9670	0.9514	0.9591
Suicidal ideation	Very high risk	0.9730	0.9730	0.9730

Source: computed by the authors from the evaluation report of the trained model.

Three points stand out. First, in the addiction proneness task the "very high proneness" class is identified perfectly. Second, errors almost always occur between neighbouring classes, and a prediction jumping from "normal" to "extremely severe" is never seen; for a screening tool with ordinal levels this error pattern is far preferable to scattered error, because the recommended level of intervention shifts by at most one step.

Third, the weakest points in the table are where the class has few samples: the "poor" class in academic motivation with an F1 of 0.6667 and the "mild" class in anxiety with 0.6154. The "mild" band is inherently the narrowest band in the official manual, and most of its errors involve the two adjacent bands.

4.5. Feature-level explainability

The explainability module performed four different computations of feature contribution for each of the eight tasks. Table 9 shows the top three features of each task according to tree Shapley values, the method that gave the most stable ranking in the stability experiments.

Table 9. Top three features of each task based on tree Shapley values

Task	Top three features (in order of importance)	Block of the first feature
Self-esteem	Disliking oneself; feeling like a failure; item 2 of the suicidal ideation scale	Depression
Anxiety	Suicidal thoughts (depression item); disliking oneself; item 14 of the suicidal ideation scale	Depression
Stress	"I feel I have failed in most areas of life"; crying; "I often suffer from trouble sleeping"	Addiction proneness
Family functioning	"I have a positive attitude towards myself"; "I prefer material that makes me think"; place of residence (east of the city)	Self-esteem
Depression	"I find it difficult to breathe"; "I feel I have little to be proud of"; "I feel distressed and adrift"	Anxiety
Addiction proneness	"I find it hard to be calm and at ease"; "I like difficult tasks because they test my abilities"; item 14 of the suicidal ideation scale	Stress
Academic motivation	"Sometimes I feel like swearing"; "On the whole I am satisfied with myself"; item 8 of the suicidal ideation scale	Addiction proneness
Suicidal ideation	"We make sure family members carry out their family responsibilities"; disliking oneself; "Someone with many problems has no choice but to turn to drugs"	Addiction proneness

Source: extracted by the authors from the out-of-fold explainability report of the trained model.

Two points stand out. First, in all eight tasks the top three features come from a scale other than the task's own; this is a direct consequence of the feature-masking mechanism, but the size of their contribution shows that they are not mere placeholders. Second, items of the suicidal ideation scale appear among the top three features in five of the eight tasks, suggesting that this scale acts largely as a general indicator of psychological distress.

Figures 4 and 5 show the top twenty features for depression and suicidal ideation. Comparing the two reveals the different viewpoints of the two model families: the tree model concentrates more on symptom items, while the deep model gives more weight to family and contextual items.

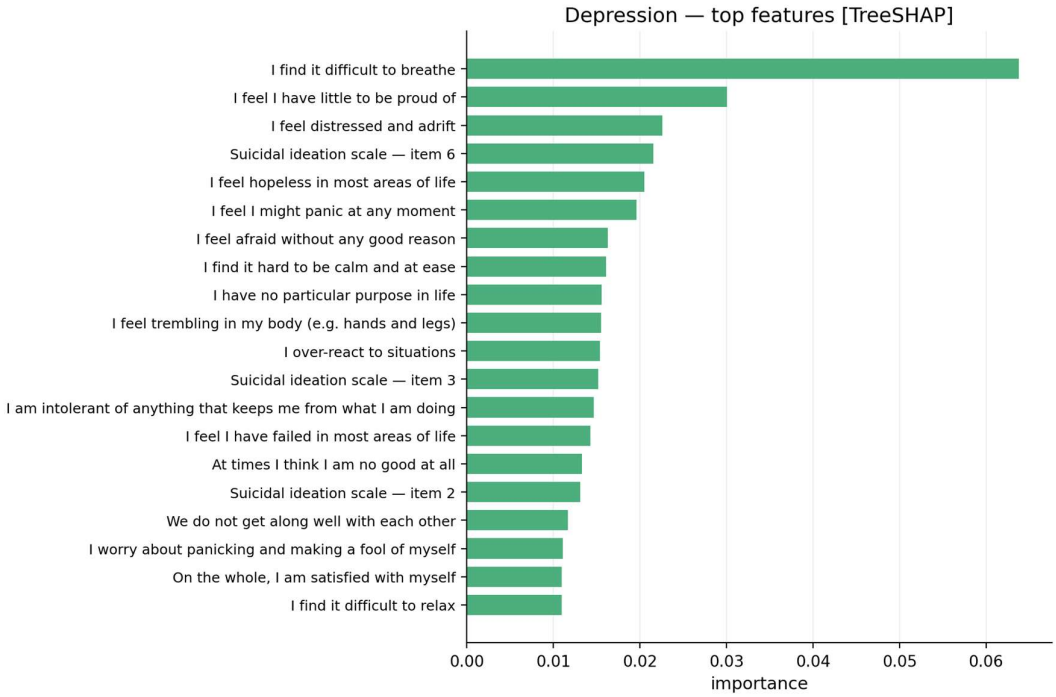

Figure 4. Top twenty features for the depression task based on tree Shapley values

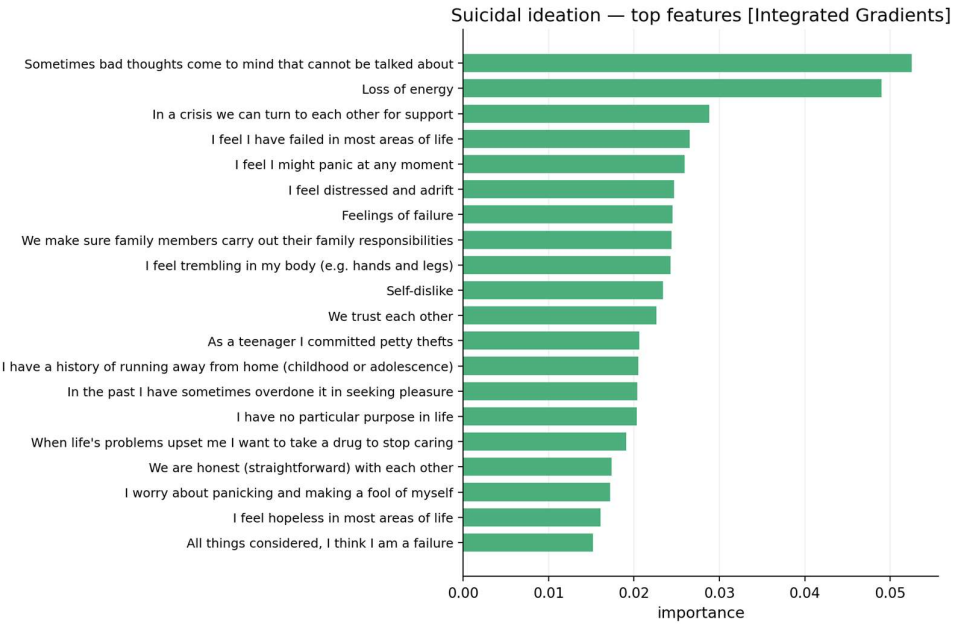

Figure 5. Top twenty features for the suicidal ideation task based on integrated gradients

To better understand the linear relationships between the features and the target variables, a correlation heat map was drawn for the thirty features with the highest mean absolute correlation with the eight targets. The map shows that:

- Features related to suicidal ideation and depression have the highest correlations with one another and with the other targets.
- Contextual variables (such as gender and economic status) correlate more weakly with the targets.
- The pattern of correlation among the features agrees with what was observed in the explainability analysis (Table 9 and Figures 4 and 5), showing that the model's non-linear relationships confirm the same patterns.

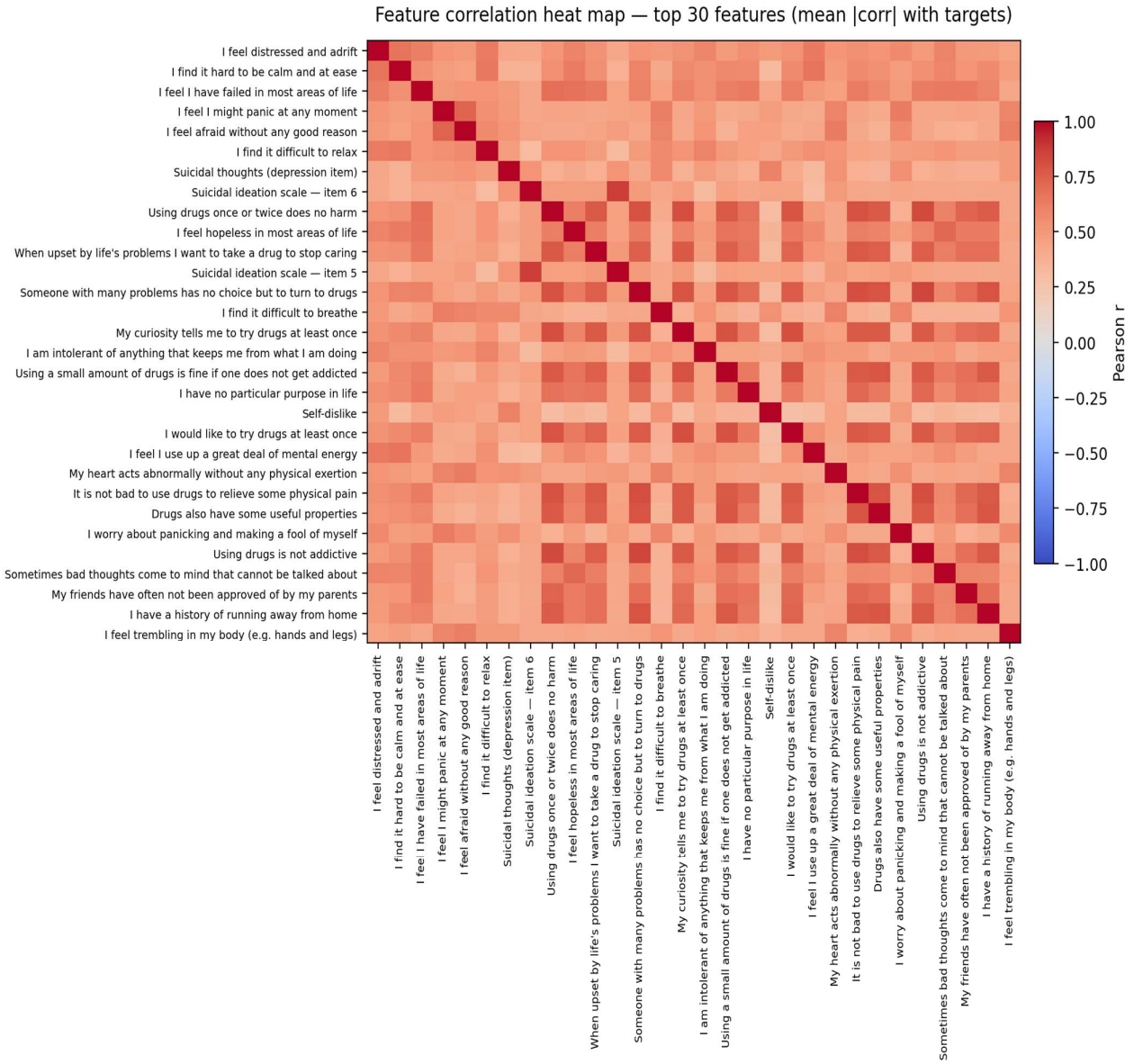

Figure 6. Correlation heat map of the top thirty features (by mean absolute correlation with the eight targets)
Source: output of the study's correlation analysis module.

4.6. *Explainability at the scale-block level*

Relying on the effect of each feature alone can be misleading, because correlated items within a scale mask one another's contribution. To address this, the contributions of the features in each block were summed. Table 10 shows the three most influential blocks for each task, and Figure 7 shows the full map of relationships among the scales.

Table 10. The three scale blocks contributing most to the prediction of each task

Target task	Most influential block	Second block	Third block
Self-esteem	Depression (39.7%)	Addiction proneness (24.0%)	Academic motivation (14.7%)
Anxiety	Depression (63.1%)	Suicidal ideation (15.6%)	Addiction proneness (15.2%)
Stress	Addiction proneness (55.3%)	Depression (36.8%)	Family functioning (4.6%)
Family functioning	Academic motivation (30.0%)	Self-esteem (19.8%)	Addiction proneness (18.7%)
Depression	Anxiety (33.8%)	Stress (21.4%)	Self-esteem (14.6%)
Addiction proneness	Suicidal ideation (34.2%)	Stress (25.0%)	Academic motivation (19.0%)
Academic motivation	Addiction proneness (41.9%)	Family functioning (16.7%)	Self-esteem (11.5%)
Suicidal ideation	Addiction proneness (37.4%)	Depression (26.7%)	Family functioning (21.2%)

Source: computed by the authors by aggregating feature contributions at the scale-block level.

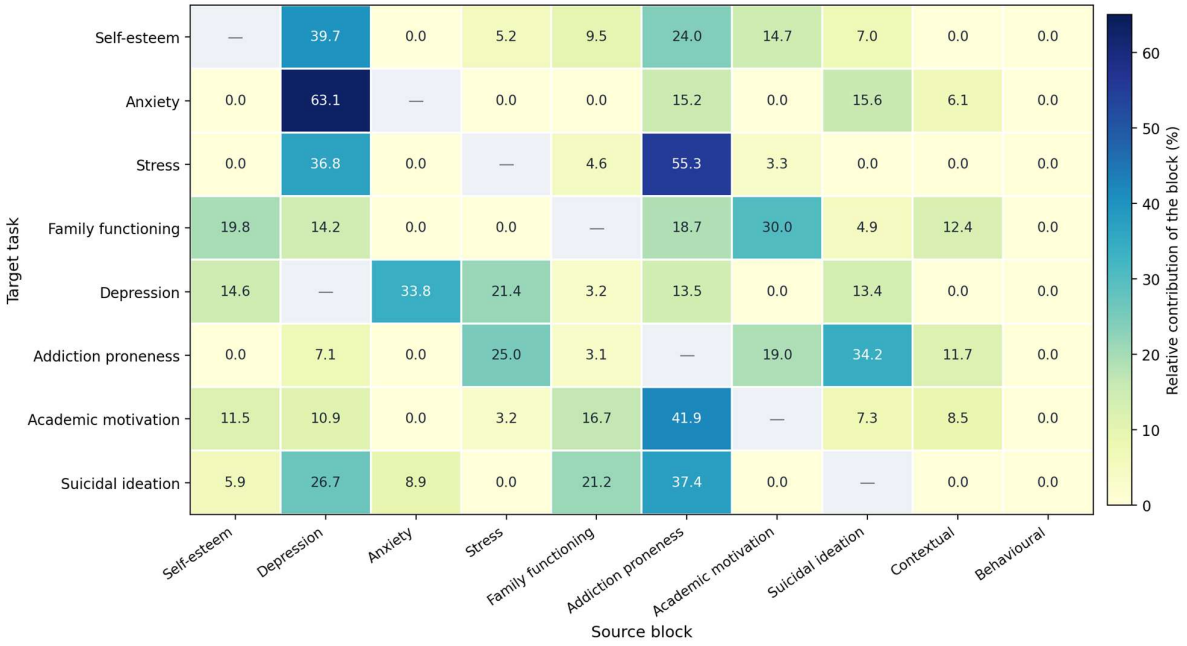

Figure 7. Cross-scale dependency map: relative contribution of each source block to the prediction of each target task
Source: drawn by the authors from the output of the explainability module.

This analysis gives a clearer picture of the role of each block. In all eight tasks the most influential block is always one of the other scales, and the contextual block never ranks first. Three meaningful patterns emerge from this map.

The first pattern is the two-way relationship between addiction proneness and suicidal ideation: the suicidal ideation block accounts for 34.2 per cent of the influence in predicting addiction proneness, and the addiction block for 37.4 per cent of the influence in predicting suicidal ideation. This same relationship had previously been observed on the same data with a completely different method and labelling scheme; its persistence indicates that it is a genuine property of the data, not the incidental result of one particular model.

The second pattern is the central role of depression: the depression block is the most influential source for anxiety, with 63.1 per cent, and for self-esteem, with 39.7 per cent, and ranks second for stress with 36.8 per cent. This can be read as a sign of a shared negative-affect core among these constructs.

The third pattern is the relative separation of family functioning and academic motivation from that affective core: the most influential source for predicting family functioning is the academic motivation block at 30.0 per cent, and the most influential source for predicting academic motivation is the addiction proneness block at 41.9 per cent. Unlike the four tasks of the affective core, neither has depression as its primary source.

4.7. Local explanation and learning curve

At the level of individual samples, the module reports not only the final probability but also the separate opinion of each base model. On confident samples all models agree; on borderline samples they diverge — in anxiety, for example, the deep model may choose "moderate" while the tree

models choose "severe" — and this disagreement can serve as a signal of ambiguity and a basis for referral to human judgement.

Figure 8 shows the learning curve for the depression task as a function of the number of training samples. The training curve stays almost flat at one, while the validation curve rises from about 0.69 on the smallest subset to nearly 0.88. The gap between the two curves narrows steadily and the slope of the validation curve is still positive at the last point, meaning that the model has not yet reached its ceiling and that adding real data could improve performance further.

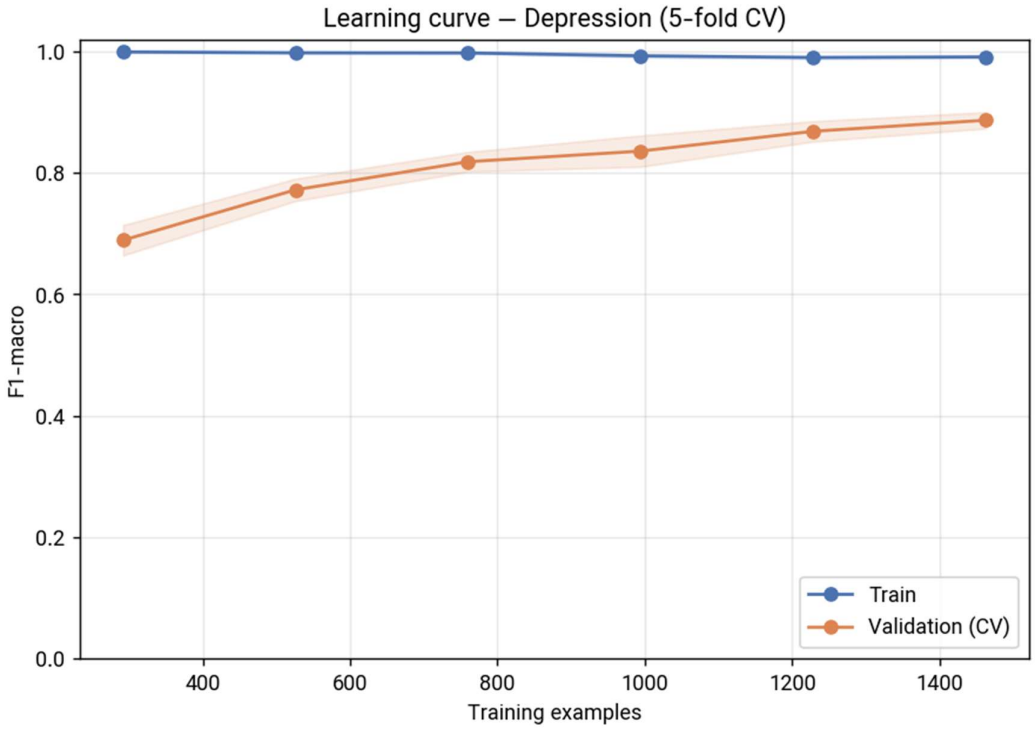

Figure 8. Learning curve for the depression task as a function of the number of training samples

Source: output of the study's training module.

4.8. Ablation Study on Model Components

To isolate the individual contribution of each architectural and methodological component to the overall performance, an ablation study was conducted on the multi-task targets using identical stratified 5-fold cross-validation splits. As detailed in Table 11, introducing the out-of-domain pre-training phase provided a substantial foundational gain (+6.33% in Macro F1), while the multi-task tabular transformer with shared attention improved joint cross-target representations (+4.35% F1). Further incorporating heterogeneous gradient-boosted trees and the meta-learning stacking layer with threshold calibration yielded the finalized macro F1 of 0.9034 and macro accuracy of 0.9277.

Table 11: Component-wise Ablation Analysis

Model Configuration / Component	Macro Accuracy	Macro F1	Macro ROC-AUC
Baseline (Homogeneous Tree Models, No Pre-training)	0.8140	0.7612	0.8845
+ Out-of-Domain Pre-training (Branch 1)	0.8690	0.8245	0.9320
+ Multi-Task Tabular Transformer (Branch 2)	0.8955	0.8680	0.9580
+ Per-Target GBT Base Learners (Branch 3)	0.9120	0.8872	0.9715
Full STEM-Net (Heterogeneous Stacking + Calibration)	0.9277	0.9034	0.9844

5. Discussion

The first question concerned the effect of pre-training on public data and of diversity in the base layer, and the results support it: the mean macro F1 is 0.9034 and the framework outperforms the baseline in all eight tasks, and does so with five more classes. Importantly, the improvement is not confined to the tasks whose instruments appeared in the public data; self-esteem, anxiety and stress were the only tasks with comparable items, yet the largest gains occurred in academic motivation and family functioning. The reason lies in the model's structure: in a tabular transformer only the feature-to-token embedding is feature-specific, and the attention blocks operate on tokens, so the body's knowledge of the general structure of Likert responding transfers to all targets.

The striking improvement in academic motivation and family functioning cannot be attributed to pre-training alone; their weakness in the baseline came from the very small number of samples in some classes. Three factors contributed simultaneously: labelling taken directly from the official manuals, which removed artificial boundaries; the presence of the pre-trained model in the base layer; and the automatic choice between stacking and simple averaging.

The second question concerned that automatic choice. The comparison showed that the winning method depends on the task and that no single method is best for all — a finding consistent with what is already known about stacked generalisation [10]: stacking works well when the meta-model has enough samples, and overfits in tasks with rare classes. Delegating this decision to an automatic comparison costs very little computationally.

The third question concerned threshold tuning: its average effect is positive but small, and more important than its size is its direction — precision rises and recall falls. This trade-off is acceptable for most tasks but not for suicidal ideation, where the cost of missing a genuine case far exceeds that of a false alarm. The practical conclusion is clear: threshold tuning should be decided per task on the basis of the cost of the errors.

The fourth question concerned the stability of the pattern of relationships among the scales, and the answer is positive. The two-way relationship between addiction proneness and suicidal ideation, previously observed on the same data with a different method and labelling scheme, holds in STEM-Net with contributions of 37.4 and 34.2 per cent. The persistence of this pattern across a change of model architecture and label levels is fairly strong evidence that it is real, and it is consistent with the evidence for the co-occurrence of these two problems in adolescence [4].

Clinical and Practical Considerations: STEM-Net is explicitly designed as a first-line screening and risk-prioritization tool for educational environments, rather than a standalone diagnostic system. Given the critical sensitivity of psychological constructs—particularly suicidal ideation—the model's predictions must not replace clinical judgment. While high recall is achieved, false negatives remain an inherent risk; therefore, any positive or borderline risk indication should immediately activate standardized human-in-the-loop referral pathways to licensed psychologists. Furthermore, explainability outputs (TreeSHAP, Integrated Gradients, and Attention Readouts) reflect statistical feature attributions and co-occurrence patterns within the observed cohort, and should not be interpreted as definitive causal relationships.

5.1. Theoretical and practical implications

Theoretically, the findings show that transfer learning on psychometric data is possible in two ways: horizontally, by sharing the feature space and the representation trunk across tasks; and vertically, by transferring from a large public set to a small local sample. The second route has a considerable effect even with little item overlap — here 24 items out of 217 features — provided that the model structure permits the transfer.

Methodologically, three points are worth noting. First, standardising the number of clinical levels is not a harmless choice: preserving the official levels not only did not reduce accuracy but, with a suitable architecture, delivered both finer granularity and higher accuracy. Second, the decision about the combination method is better left to the data itself. Third, reporting feature contributions without showing the margin of error can be misleading.

Practically, the most important advantage of STEM-Net is that its output levels are exactly the levels of the instruments' official manuals; the school counsellor does not need to translate the model's output, and "severe anxiety" in the model's output carries the same meaning as in the official manual. In addition, disagreement among the base models provides a natural, low-cost signal for flagging doubtful cases and referring them to human judgement.

5.2. *Operational deployment: the web-based screening system*

A screening tool that remains research code is of no use in a school; STEM-Net was therefore built as a web-based system with a Persian, right-to-left interface. The main obstacle is the sheer number of questions: completing all eight questionnaires takes 192 items, and respondent fatigue reduces both the number of complete responses and their quality.

The system's solution rests on the output of the explainability component: the four-method analysis had shown that, for each of the eight domains, only a limited number of features has any real effect. A short questionnaire was built from these features, asking only the contextual variables and those key items; the output of the explainability module thus feeds directly into the design of the data-collection instrument.

This shortening is only possible because the feature-masking mechanism forced the model from the outset to predict each domain from the items of the other scales; the very restriction imposed to protect label validity turns, at deployment, into a practical advantage.

The system's output is shown separately for each of the eight domains and at several levels. Importantly, explainability is applied here at the individual level: for each student the system identifies which features most influenced the model's decision about that student and displays them alongside the predicted level and a matching recommendation. The counsellor therefore receives not just a label but a view of where that label came from.

Two safety measures are also built into the system. First, whenever the predicted level of suicidal ideation rises above the lowest level, an immediate referral message and crisis support resources are displayed prominently. Second, the system states throughout that its output is an educational and research screening tool and not a substitute for clinical diagnosis or treatment. A user feedback facility is also provided.

5.3. *Limitations*

First, the data were collected in a single city using convenience sampling, and generalising the results to other populations requires separate investigation. Second, all experiments were run with a single random seed and the reported figures carry no confidence intervals; repeating them with several seeds would give a more accurate estimate of the stability of the results.

Ethically, the output of this framework is under no circumstances a clinical diagnosis and must not be used as a basis for labelling a student. Despite a recall of 96.81 per cent on the suicidal ideation task, some high-risk cases still go undetected; using it without expert human supervision and without a referral support mechanism is not defensible.

A pertinent methodological consideration is whether deploying a multi-branch framework—incorporating a multi-task tabular transformer, heterogeneous gradient-boosted trees, and stacked meta-learners—induces over-engineering (given our cohort size, $N = 1{,}713$). We explicitly addressed this vulnerability through four complementary regularization safeguards: (1) leveraging out-of-domain pre-training (TabPFN-style initialization) to provide stable inductive priors prior to local fine-tuning; (2) enforcing aggressive stochastic regularization, including feature dropout and attention head weight decay; (3) employing strictly out-of-fold stratified 5-fold cross-validation for meta-learner training, preventing target leakage across ensemble layers; and (4) empirical verification via our ablation analysis (Table 11), which demonstrates consistent monotonic generalization gains across independent holdout validation splits rather than divergence between training and validation trajectories. Consequently, architectural capacity was directly harnessed for representation richness without incurring sample-size-driven overfitting.

6. Conclusion

This study introduced and evaluated STEM-Net, a framework for the simultaneous prediction of eight psychological and academic problems in high-school students. Three methodological pillars underpin the work: transfer pre-training on nearly eighty-eight thousand public respondents; a base layer of five heterogeneous learners in three branches, together with an automatic per-target choice of combination strategy; and labelling taken directly from the official interpretation of the instruments, which turns the problem into a twenty-nine-class ordinal classification.

On the test set the resulting framework reached a macro-average accuracy of 0.9277, an F1 of 0.9034 and a ROC-AUC of 0.9844. The comparison with SAFE-Net showed that STEM-Net performs better in all eight tasks, and does so on a harder distinction with 29 classes rather than 24; output granularity and accuracy therefore rose together, and

the largest gains occurred in the two tasks where the baseline was weakest. The pattern of cross-scale dependency — in particular the symmetry between addiction proneness and suicidal ideation and the central role of depression — also remained stable.

On the basis of these findings, three directions are proposed for future work: a study that separates the contribution of each of the three methodological changes, since all three were applied together; extending the pre-training data to more instruments so that the item overlap grows beyond the current 24 items; and repeating the experiments with several random seeds and reporting confidence intervals.

Beyond the laboratory evaluation, the framework was deployed as a web-based screening system; the influential features identified in the explainability analysis formed the basis of a short questionnaire, and for each student the system reports the predicted level in each domain together with that individual's most influential features.

At the practical level, two recommendations follow: setting a confidence threshold and automatically referring low-confidence cases to human judgement using the disagreement among the base models; and designing a system that shows the counsellor the influential items alongside the predicted level.

Authors' Contributions

All authors equally contributed to this study.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

By researchers in Bioinformatics, Computational Biology, Artificial Intelligence and Medicine.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] World Health Organization, "Mental health of adolescents [Fact sheet]," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health.

[2] R. Qasrawi, S. P. Vicuna Polo, D. Abu Al-Halawa, S. Hallaq, and Z. Abdeen, "Assessment and prediction of depression and anxiety risk factors in schoolchildren: Machine learning techniques performance analysis," *JMIR Formative Research,* vol. 6, no. 8, p. e32736, 2022, doi: 10.2196/32736.

[3] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Mining and Knowledge Discovery,* vol. 10, no. 3, p. e1355, 2020, doi: 10.1002/widm.1355.

[4] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, "Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication," *Archives of General Psychiatry,* vol. 62, no. 6, pp. 593-602, 2005, doi: 10.1001/archpsyc.62.6.593.

[5] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv,* 2017, doi: 10.48550/arXiv.1706.05098.

[6] B. Saylam and Ö. Durmaz İncel, "Multitask learning for mental health: Depression, anxiety, stress (DAS) using wearables," *Diagnostics,* vol. 14, no. 5, p. 501, 2024, doi: 10.3390/diagnostics14050501.

[7] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 35, no. 6, pp. 7499-7519, 2024, doi: 10.1109/TNNLS.2022.3229161.

[8] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," 2021, vol. 34, pp. 18932-18943, doi: 10.48550/arXiv.2106.11959.

[9] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," in *International Conference on Learning Representations*, 2023, doi: 10.48550/arXiv.2207.01848.

[10] D. H. Wolpert, "Stacked generalization," *Neural Networks,* vol. 5, no. 2, pp. 241-259, 1992, doi: 10.1016/S0893-6080(05)80023-1.

[11] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," 2017, vol. 30, pp. 4765-4774. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[12] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International*

Conference on Machine Learning, 2017, vol. 70, pp. 3319-3328. [Online]. Available: http://proceedings.mlr.press/v70/sundararajan17a.html.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[14] M. Rosenberg, *Society and the adolescent self-image*. Princeton University Press, 1965.

[15] S. H. Lovibond and P. F. Lovibond, *Manual for the Depression Anxiety Stress Scales*, 2nd ed. Psychology Foundation of Australia, 1995.

[16] A. T. Beck, R. A. Steer, and G. K. Brown, *Manual for the Beck Depression Inventory-II*. Psychological Corporation, 1996.

[17] N. B. Epstein, L. M. Baldwin, and D. S. Bishop, "The McMaster Family Assessment Device," *Journal of Marital and Family Therapy,* vol. 9, no. 2, pp. 171-180, 1983, doi: 10.1111/j.1752-0606.1983.tb01497.x.

[18] A. T. Beck, M. Kovacs, and A. Weissman, "Assessment of suicidal intention: The Scale for Suicide Ideation," *Journal of Consulting and Clinical Psychology,* vol. 47, no. 2, pp. 343-352, 1979, doi: 10.1037/0022-006X.47.2.343.

[19] Y. Zargar, "Construction and validation of the Iranian Addiction Potential Scale," 2006.

[20] 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, doi: 10.1613/jair.953.

[21] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011. [Online]. Available: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page.