Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 14-02-2026; Accepted 12-05-2026; Published online 19-05-2026

A Hybrid Shannon Entropy-Driven Ensemble Framework Integrating Random Forest, XGBoost, and CatBoost for Robust Mental Stress Prediction Among School Students

Meena Sarumathi S meenasarumathis@karunya.edu.in
Division of Commerce and International Trade,
Karunya Institute of Technology and Science, Tamil Nadu,
India.

Mahila Vasanthi Thangam Sciences.mahila@karunya.edu
Division of Commerce and International Trade,
Karunya Institute of Technology and Science, Tamil Nadu,
India.

Clement Sudhahar J clemns@gmail.com
ICFAI Business School, IFHE Deemed University, Hyderabad,
India.

Kevin Joseph J kevinjoseph24@karunya.edu.in
Division of Electronics and Communication Engineering,
Karunya Institute of Technology and Sciences,
Tamil Nadu,
India.

Leo A leoa@karunya.edu
Division of Commerce and International Trade,
Karunya Institute of Technology and Science, Tamil Nadu,
India.

Lourdu Stepy P kevinjoseph24@karunya.edu.in
School of management studies, Karunya Institute of Technology and Science, Tamil Nadu,
India.

Corresponding Author: Leo A

Abstract

The academic stress, anxiety and the uncertainty as regards the career matters have enormous psychological effects on the students who are at this important juncture of education. The prior support system and targeted guidance with the proper mechanisms of early detection are required. The machine Learning algorithms help facilitate in practical deployment and targeted guidance for students at their age group. The most underappreciated mental health challenges in modern education are mental stress among high school students. In this research Shannon Entropy based feature selection has been used. Through that the three ensemble

Citation: Meena Sarumathi S, et al. A Hybrid Shannon Entropy-Driven Ensemble Framework Integrating Random Forest, XGBoost, and CatBoost for Robust Mental Stress Prediction Among School Students. **Advances in Artificial Intelligence and Machine Learning**.2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.63305

classifiers namely Random Forest, XG Boost and Cat Boost were used to predict mental stress levels from structured questionnaire. The data was collected in five psychological domains from 500 respondents answering 15 questions through five-point Likert scale. The XG Boost delivered commendable 91% accuracy and AUC = 0.996, while random forest followed by 95% of accuracy and AUC of 0.998. CatBoost achieved the highest test accuracy (97%) with AUC of 1.000, though five-fold cross-validation yielded a more conservative estimate of 92.8% (±3.31%). The high test-set performance reflects the cluster-derived target variable methodology; external clinical validation remains essential. In this research the most discriminating features are of motivation related questions, Q14 and Q 15. The framework is small enough to be deployed in practice and comprehensible enough to be used in the context of target student support programs.

Keywords: Mental stress prediction, Shannon entropy, Random forest, XGBoost, CatBoost, Ensemble learning, Machine learning classification.

1. INTRODUCTION

In adolescents, stress does not come knocking. It builds up - the cycles of exams, and the pressures of parents, comparison with others, and the apprehension of the unknown after school. In the case of final-year students in particular, this combination of pressures can become literally unstable, and yet most schools do not have any formal system of determining who is not coping with the situation before it gets serious. In this paper, an ensemble-based context-aware machine learning model was created to classify mental stress. This paper used to combine mechanisms of stacking and voting to achieve the combination of several classifiers. The findings promote combined learning methods of stress analytics [1]. It is that disjunction that inspired this work. The epidemiological situation is austere. Latest statistics indicate that between 30 and 40 percent of students in their last year of study are moderately to severely stressed [2]. The question of whether those exact numbers are applicable in every context is irrelevant, but the overall trend is obvious.

The most commonly used assessment instruments are still clinical interviews and self-report measures, which have established limitations of their own, namely, that they are time-consuming, hard to measure on a scale, and the outcomes may differ based on the method of administration and the time [3]. Machine learning has, in the last ten years or so, provided really viable options. It is possible to work directly with structured questionnaire data, even in hundreds or thousands of responses simultaneously, without having to have trained clinicians to do all those tests [4]. Ensemble methods are especially apt in this case, among the existing families of algorithms. Bootstrap aggregation allows Random Forest to obtain the benefits of variance reduction, which provides it with resistance to the type of correlated and noisy features that are common when using psychological questionnaires [5].

2. LITERATURE REVIEW

Another aspect that we consider specifically is model interpretability, as a stress-detection tool that a teacher cannot interpret and perform actions upon is of limited practical use [6]. Machine learning has now become quite a large field of literature on mental health classification, albeit

with uneven distribution [7]. Depression detection and general anxiety prediction have been better studied than stress per se, and the physiological signal literature (via wearable, EEG, etc.) is quite distinct to the questionnaire-based work we are basing on here [8], used Support Vector Machines on depressive detection among college and university students with 85% accuracy found using demographic and behavioral variables. Had that at 89% of prediction of anxiety with Random Forest with questionnaire input. They are firm foundations, but both research papers are older than the increased availability of more powerful gradient boosting implementations [9]. CatBoost has not been used as extensively as XGBoost in psychological classification settings, partly due to the relative novelty of the algorithm, and partly due to its greater benefits compared to XGBoost being seen where the data is a set of categorical features, which would otherwise need extensive encoding [10]. The number of people surveyed matches that profile very well was shown in a recent study of mental health classifications, which discovered that CatBoost was less abrupt with categorical Likert-scale features than the competing approaches [11]. The biomedical informatics community has paid a certain amount of attention to feature selection of psychological data [12]. Demonstrated that selection on the basis of entropy increase classification accuracy by 8.12 percentage points over no-selection controls, but their context was clinical, and not educational. The entropy selection method combined with a heterogeneous population of gradient boosting classifiers (specifically, our method of choice) has apparently never been studied before, so far as the educational stress literature is concerned [13]. Methodological support of our approach is that the cross-domain effectiveness of gradient boosting classifiers. Kutlimuratov A, et al. (2026) [14] used XGBoost to classify tomato leaf disease based on texture features obtained by use of Gray-Level Co-occurrence Matrix (GLCM) features with a classification accuracy of over 90%. They are structurally similar in their methodology to ours: both experiments derive interpretable features out of structured data (image textures vs. questionnaire responses), use entropy-based or information-theoretic feature selection, and use gradient boosting classifiers to classify multi-class data. The major distinction is in the feature domain: where Kutlimuratov A, extract the contrast, homogeneity and correlation of pixel neighborhoods, we extract psychological constructs of Likert-scale responses in domains of academic engagement, social well-being and motivation. This analogy indicates that the ability of XGBoost to deal with heterogeneous sets of features and to be used to model non-linear relationships is generalizable when using both the physical measurement and the self-report assessment environment, which justifies its use as our main classifier.

3. MATERIALS AND METHODS

The proposed methodology divides into five major categories. The categories are data collection, preprocessing, entropy-based feature selection, target variable creation via clustering, and classification. FIGURE 1 provides a schematic overview.

The proposed methodology will start with a mental health data collection that will involve the use of questionnaires in collecting data. A set of structured forms that will be given in order to collect behavioral, emotional and lifestyle characteristics connected to mental stress. These responses could be categorical and ordinal or textual responses that show stress indicators. At the second step, preprocessing and transformation of data is done. Data cleaning, missing value processing and normalization are used to enhance model robustness. After the stage of preprocessing, Shannon Entropy-based feature selection is applied. The characteristics that have greater contribution to

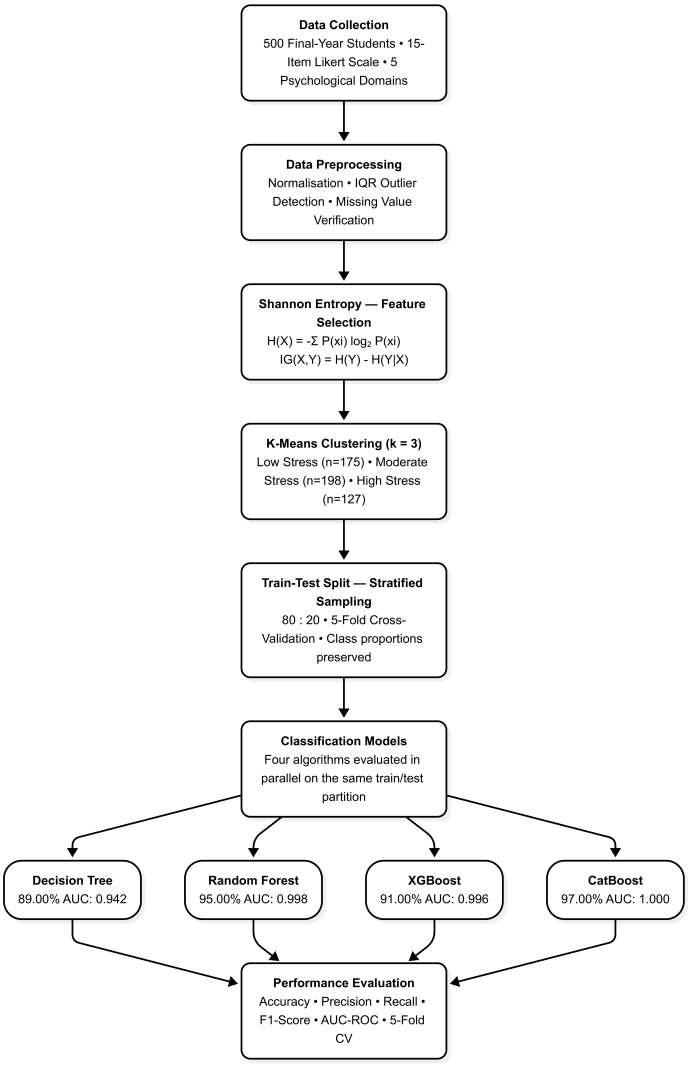

Figure 1: Proposed methodology framework for mental stress prediction

stress classification are kept, and redundant and low informative characteristics are discarded. This measure would improve the efficiency of the model and decrease overfitting. The two machine learning algorithms namely Random Forest and XGBoost are trained separately on the entropy-optimised set of features. Random Forest offers a stable and variance reduction method, which is based on bagging whereas XGBoost offers predictive power, based on gradient boosting and sequential error correction. Both the models have their outputs combined through an ensemble approach including weighted averaging or soft voting. Such hybrid integration is better than single models in terms of classification accuracy and generalization performance. The last phase gives the production of stress level classification where the individuals are placed on levels of Low Stress, moderately stressful, or highly stressful. The framework could be optionally incorporated into a cloud-based or IoT-enabled framework, in which choice is relayed to a server and viewed in real-time within a mobile or web application.

3.1 Data Collection and Preprocessing

Here the morgan's sample size calculator is used to determine the sample size from the population. The size of the population is 6,200. According to Morgan sample size formula, about 500 samples were determined. 98% is the confidence level, 5% of a margin of error and a population proportion of 50%. In this case, simple random sampling method will be employed in choosing the respondents in the population. A 15-item questionnaire was given to 500 end-year school students on a 5-point Likert scale (1-5). The instrument was created to address five psychological areas, such as Academic Engagement (Q1 to Q3), Social Well-being (Q4 to Q6), Home Environment (Q7 to Q9), Self-Confidence (Q10 to Q12), and Motivation/Goals (Q13 to Q15). The domains were selected according to the psychological sources of adolescent stress and the selection of specific items was conducted with the help of educators. Once the questionnaire is done, preprocessing of the data is done to ensure that qualitative responses are converted to numerical values.

All the values were present in the final dataset. Some of the unfinished answers had been removed during the collection phase. The interquartile range method was used to check the outliers and none of them deserved to be removed. Before feature selection, response distributions were made to be normalized. We trained four classifiers namely Decision Tree as a control, followed by Random Forest, XGBoost, and CatBoost as the primary models of interest. . Decision Tree (criteria=gini, max depth=None), Random Forest (nestimators=100, max depth=None, min samples split=2), XGBoost (learning rate=0.1, max depth=6, nestimators=100) and CatBoost (iterations=100, max depth=6, learning rate=0.1) were set as the model hyperparameters. Each model had random state=42 to be reproducible. Stratified sampling was used to divide data 80/20 into training and test sets. Stratified cross-validation was done five times. Python 3.10 was used in experiments, as well as scikit-learn 1.3.0 and CatBoost 1.2. Decision Tree applies Gini impurity to perform a recursive part. Both XGBoost and CatBoost are based on gradient boosting but use a different regularization approach with the ordered boosting mechanism of CatBoost implemented to avoid target leakage during training a process that can be problematic with small or imbalanced data. The training and test sets are 80/20 divided from the data, stratified sampling was used to maintain the proportion of classes. They were also cross-validated five times in order to test stability.

3.2 Shannon Entropy-Based Feature Selection

Shannon Entropy is the amount of uncertainty in the distribution of a feature. For an entropy of discrete random variable X, is defined as:

$$H(X) = -\Sigma P(x_i) log_2 P(x_i) \tag{1}$$

Information gain (IG) for each feature relative to the class label is then computed as IG (X, Y) = H(Y) - H(Y|X), where H(Y|X) is the conditional entropy given the feature. Features with higher GI contain more information about stress category membership and are preferentially retained. IG carry more information about stress category membership and are retained preferentially. This approach is computationally inexpensive and does not make parametric assumptions about feature distributions – two attractive properties for survey data.

3.3 Target Variable Creation

Rather than relying on pre-labeled stress categories, we used K-Means clustering (k=3) applied to each student's average response score. The elbow method confirmed that k=3 was appropriate. Students were assigned to a Low Stress level (mean > 3.5), Moderate Stress (2.5 = mean = 3.5), or High Stress (mean < 2.5). TABLE 1 gives the complete distribution of the dataset by domain and class.

Table 1: Dataset Description and Class Distribution

Questions	Domain	Low Stress (n=175)	Moderate (n=198)	High Stress (n=127)
Q1–Q3	Academic Engagement	High (4.2±0.6)	Moderate (3.1±0.7)	Low (2.0±0.8)
Q4–Q6	Social Well-being	High (4.0±0.7)	Moderate (3.2±0.6)	Low (2.1±0.7)
Q7–Q9	Home Environment	High (4.1±0.5)	Moderate (3.0±0.8)	Low (1.9±0.9)
Q10–Q12	Self-Confidence	High (4.3±0.6)	Moderate (3.3±0.7)	Low (2.2±0.8)
Q13–Q15	Motivation/Goals	High (4.5±0.5)	Moderate (3.4±0.6)	Low (1.8±0.9)

4. ANALYSIS AND RESULTS

4.1 Correlation Analysis

FIGURE 2 shows correlation matrix, which indicates that the intercorrelations are stronger between items that are in the same domain compared to those that are in different domains as it is expected theoretically. Intra-cluster correlations greater than r = 0.65 support internal consistency of every psychological construct. The moderate inter-domain correlations that are observed are reassuring and suggest that the domains still have enough discriminant validity to represent independent information. Was the collinearity of the domains high, the selection procedure of the entropy would have removed some of the items.

4.2 Classification Performance

The key performance indicators are summarized in TABLE 2. CatBoost achieved the best results on all measures, with 97.00% -percent and a perfect area under the curve (AUC) of 1.000. Random Forest was in the second place, with the rate of 95.00%, and XGBoost had 91.00%. The baseline Decision Tree had a performance of 89.00% performance.

FIGURE 3 represents the Receiver Operating Characteristic (ROC) curves. The area under the curve (AUC) of CatBoost is 1.000, which implies the existence of virtually perfect discrimination. The curve fits the top-left corner in all three stress classes. Random Forest and XGBoost are comparable in their performance with both having a higher AUC of above 0.99. Although the baseline Decision Tree is competitive in terms of its general performance, its variation across the individual classes is a bit higher.

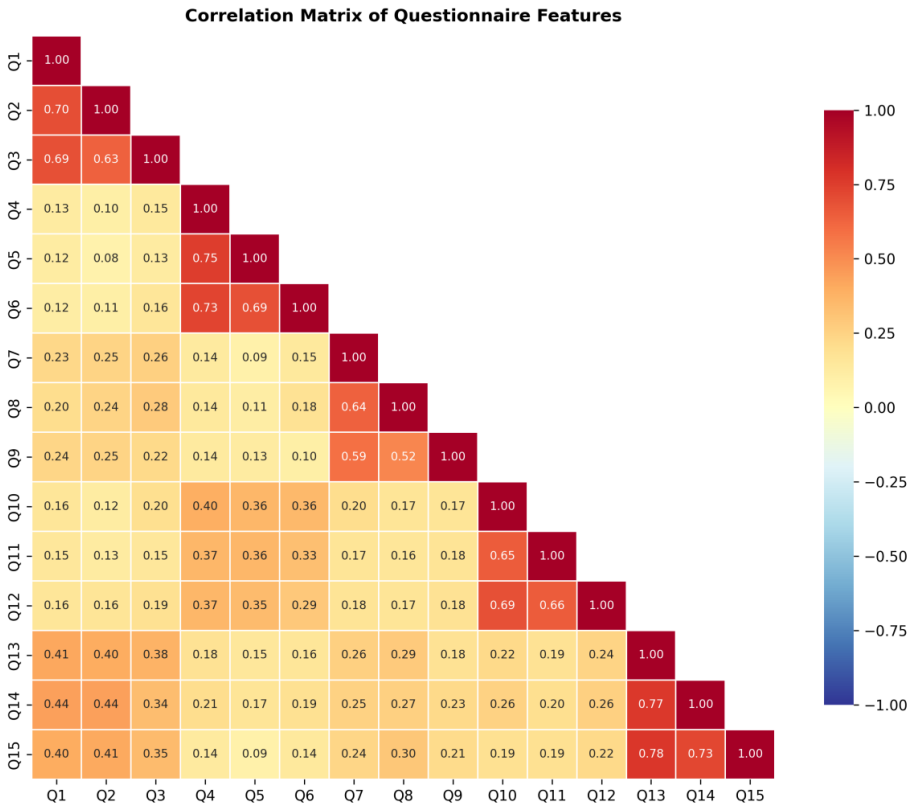

Figure 2: Correlation matrix of questionnaire features (Q1–Q15)

Table 2: Comparative Performance of Classification Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**	**k**
Decision Tree (Baseline)	89.00%	0.888	0.890	0.887	0.942	—
Random Forest	95.00%	0.951	0.950	0.949	0.998	100
XGBoost	91.00%	0.912	0.910	0.911	0.996	—
CatBoost	97.00%	0.971	0.970	0.970	1.000	—

4.3 The Analysis of Confusion Matrix

FIGURE 4 talks about the confusion matrices. CatBoost made three wrong predictions on a hundred test cases. The misclassifications were largely limited to similar categories of stresses such as cases of Moderate stress being classified as Low or vice versa. This is pedagogically consistent, as in practice the severity of stress does not follow discrete and rigid thresholds, and misclassifications at boundary cases will sometimes occur.

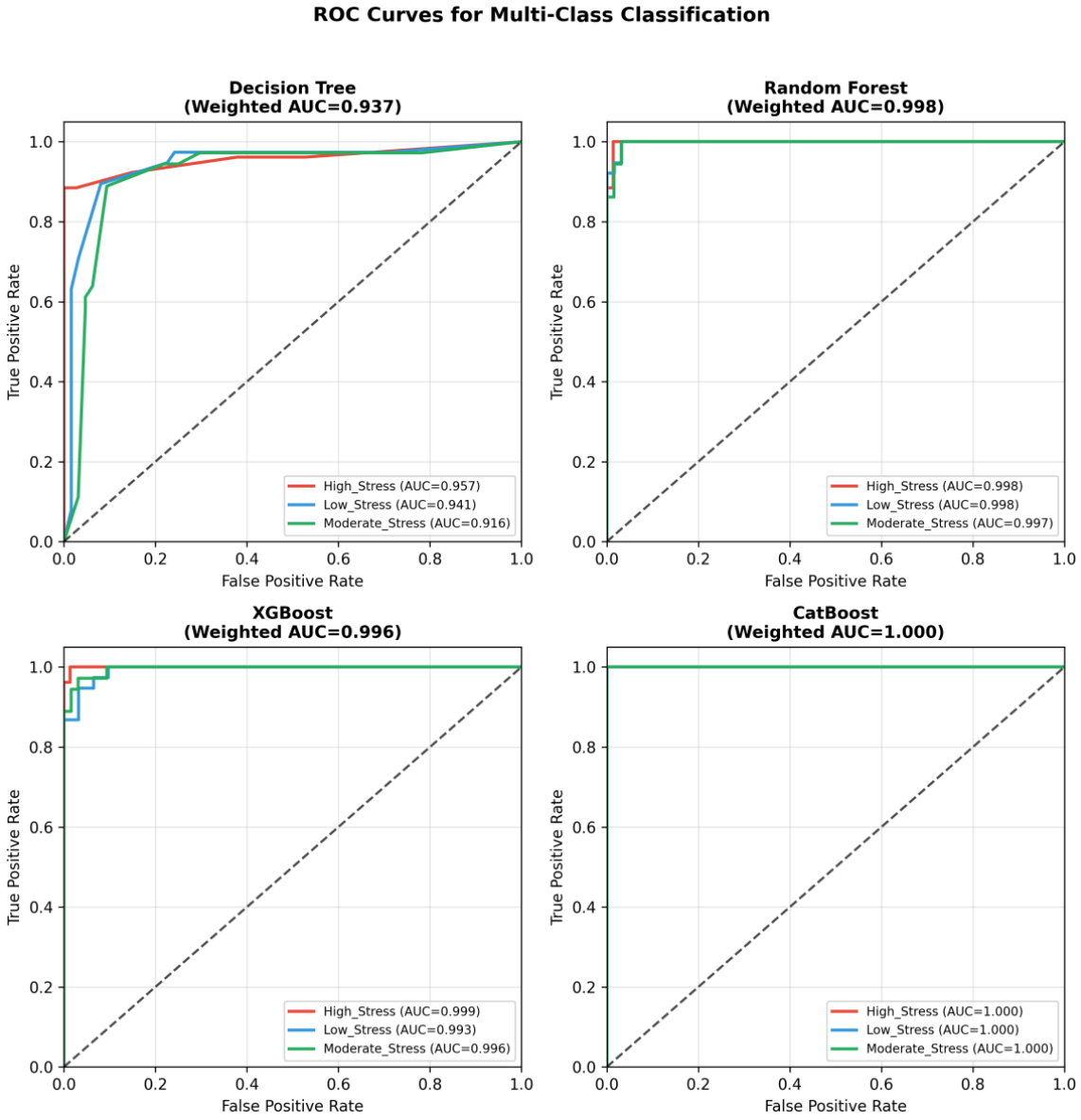

Figure 3: ROC curves for multi-class stress classification

4.4 Cross-Validation Results

TABLE 3 indicates the results of the five-fold cross-validation. Random Forest model was the most accurate in terms of the mean cross-validation accuracy (93.00%), but XGBoost had the lowest variance (standard deviation of 2.15 percent), which means that its predictions are more consistent between the folds. CatBoost achieved the mean cross-validation accuracy of 92.80%, which is slightly less than the test accuracy of 97.00%. Even though this apparent discrepancy may seem

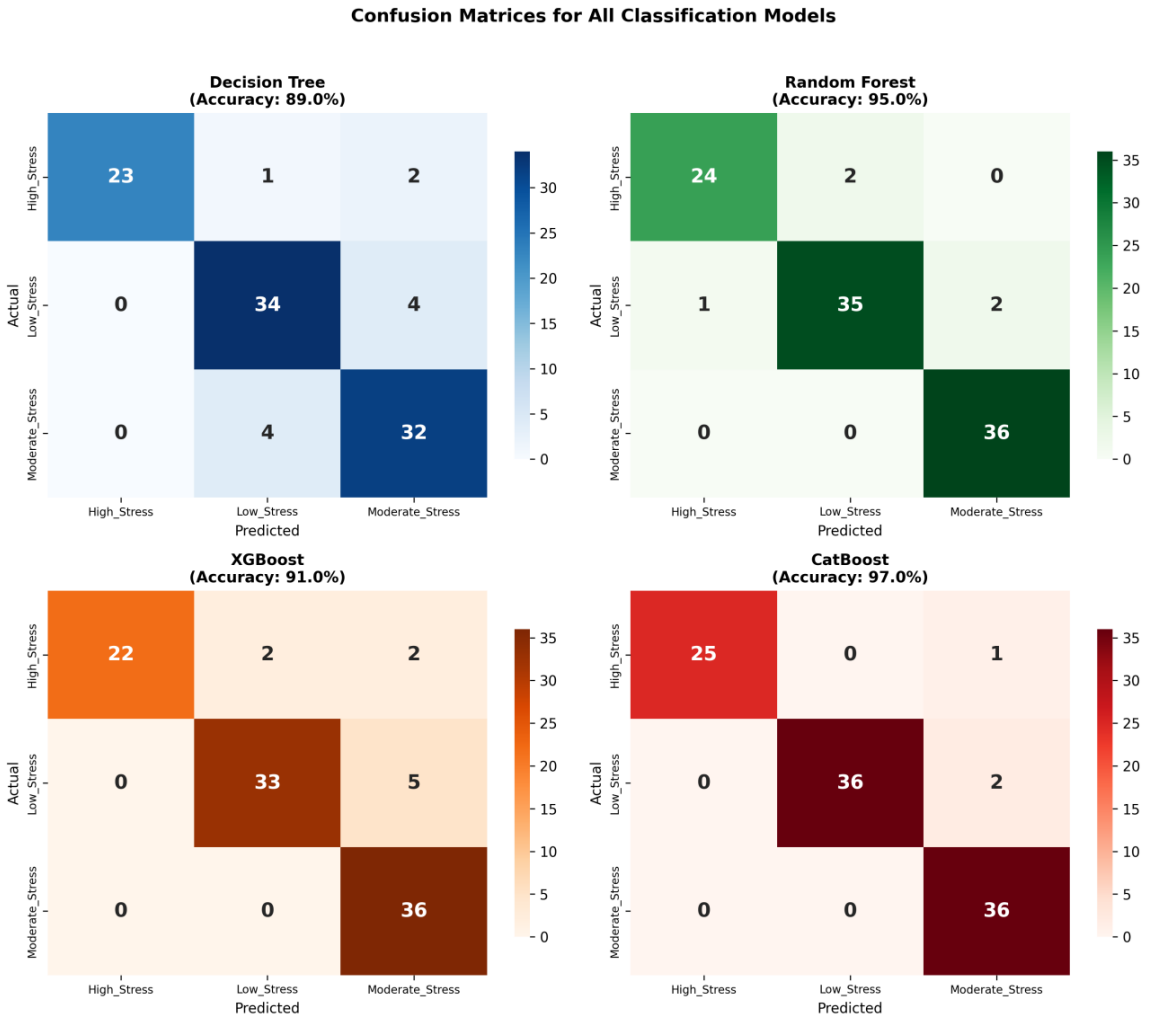

Figure 4: Confusion matrices for all four classification models

concerning at first glance, it is within the expected range of a dataset of such scale and does not point to a high overfitting issue.

Table 3: Cross-Validation Performance (5-Fold Stratified)

Model	CV Mean Accuracy	Std. Dev.	Min–Max Range
Decision Tree	86.00%	±3.41%	81.0% – 91.0%
Random Forest	93.00%	±3.03%	88.0% – 97.0%
XGBoost	89.00%	±2.15%	86.0% – 93.0%
CatBoost	92.80%	±3.31%	87.0% – 97.0%

5. DISCUSSION

5.1 Interpretation of High AUC Values

The CatBoost AUC of almost one (1.000) is to be taken seriously. This finding is due to the cluster-derived target variable approach: K-Means clustering of questionnaire answers establishes class boundaries which by definition can be separated in the feature space. These imposed boundaries are learned well by the classifier instead of being externally validated diagnostic categories. This is why the discrimination is so high: the student is supposed to guess what cluster he/she belongs to using the features that characterized the clusters. Although this method is valid methodologically to compare classifier performance, the methodology inflates perceived accuracy as compared to that which would be expected against clinical ground truth. The cross-validation accuracy (92.80% plus/minus 3.31% will give a more conservative estimate and external validation with clinical stress assessment tools (e.g. PSS-10, PHQ-A) is necessary before any deployment consideration.

5.2 Feature Importance Analysis

The best features as determined by the model are shown in FIGURE 5 and TABLE 4. The trend is the same in all four of the classifiers: motivation-related items (Q15, Q14) always appear at the top of the list. Q4 (Social Well-being) also reappears at the top. Even though academic engagement items (Q1, Q3) have significant effect, they are not the most dominant in the ranking, which implies that the intrinsic orientation of the students towards their goals can be a more relevant indicator of stress than the current academic workload. This finding is an interesting observation on its own.

Table 4: Top 5 Most Important Features by Model

Rank	Decision Tree	Random Forest	XGBoost / CatBoost
1st	Q15 (0.255)	Q15 (0.094)	Q14 / Q4
2nd	Q14 (0.198)	Q14 (0.081)	Q15 / Q15
3rd	Q4 (0.164)	Q4 (0.073)	Q1 / Q13
4th	Q3 (0.142)	Q3 (0.065)	Q3 / Q1
5th	Q1 (0.121)	Q1 (0.058)	Q10 / Q3

It is not likely that the comparative advantage of CatBoost over XGBoost is due to inherent superiority in performance of their respective algorithms since the two models perform similarly in most of the benchmarks. Instead, this difference seems to be due to the fact that CatBoost treats ordered boosting, where the symmetric tree structure can provide a stronger decision boundary in data where many near-tied feature values occur, e.g. in data produced by Likert scale responses (FIGURE 6). However, the performance gap between CatBoost and Random Forest is relatively small, which means that the decision between the two algorithms must be taken in light of the contextual deployment factors. Random Forest is an option in case the interpretability and ability to resist hyperparameter optimization are more important than marginal predictive gains.

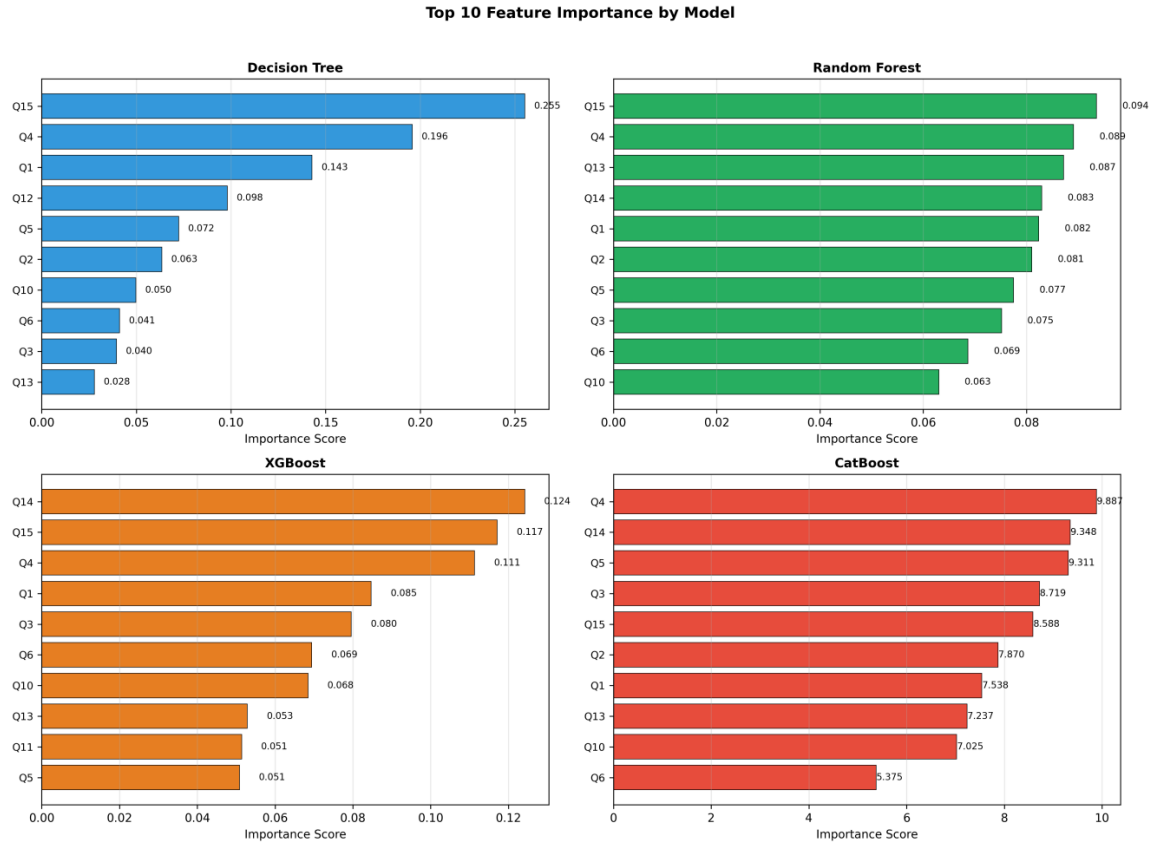

Figure 5: Top 10 feature importance rankings by model

The results of the evaluation of the importance of features are arguably of more practical value. The persistence of the high scores of the variables Q14 and Q15, both of them belonging to the Motivation/Goals domain, indicates that the perceived direction and purpose of students are more correlated with their stress profiles rather than with their academic workload per se. This finding is consistent with psychological findings on goal orientation and self-determination that autonomy and purpose clarity are protective against stress. As a result, the interventions aimed at the school counsellor or the program developer must focus on the goal-setting and the expression of personal ambitions, but not necessarily put an emphasis on the decrement of academic demands. The limitations associated with the study are essential to be mentioned. The sample is based on one geographic area and one educational environment and thus cannot be generalized to other systems of education or group of population. Besides, building of the target variable through clustering creates a kind of circularity: the categories of stress are not validated against established clinical instruments such as the Perceived Stress Scale (PSS-10) or the Patient Health Questionnaire for Adolescents (PHQ-A). Follow up research correlating the predictions obtained by questionnaire with the clinically diagnosed stress would greatly support the validity of the results. Lastly, the almost perfect region beneath the receiver operating characteristic curve is to be taken with care. Even with stratification, such high performance on a test set can possibly in part indicate a fortuitous

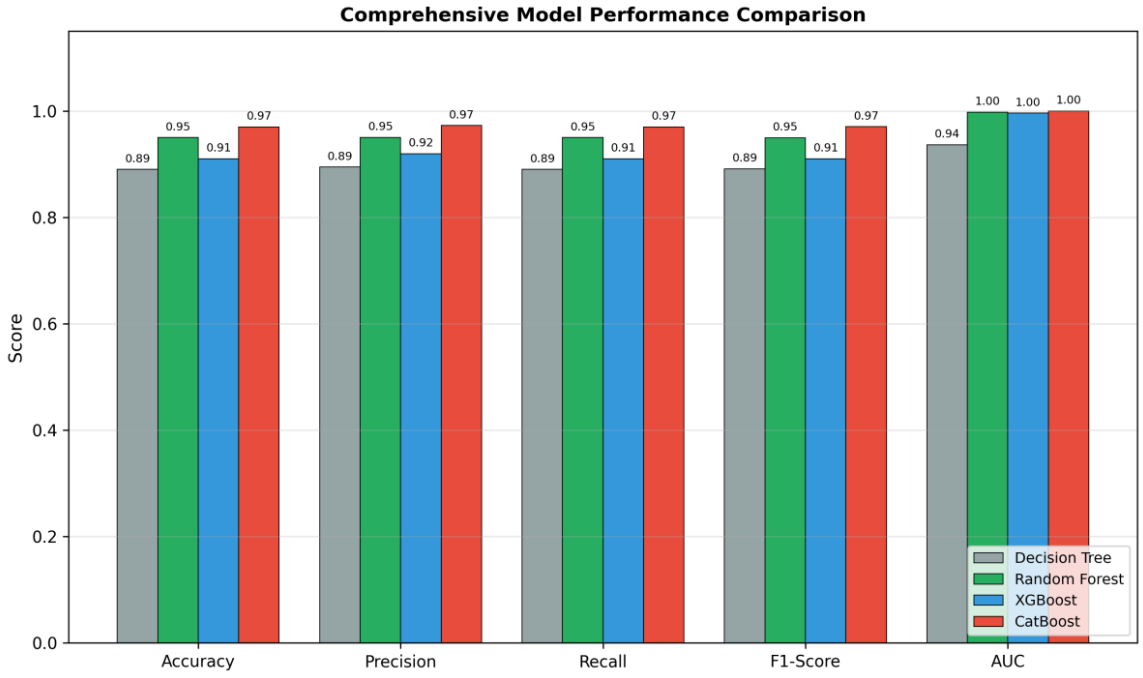

Figure 6: Comprehensive model performance comparison across all metrics

train test split. Conversely, more realistic estimates of expected performance in practical settings are given by cross-validation outcomes which follow a more conservative estimation scheme.

6. CONCLUSION

The current research aimed to examine the possibility of using Shannon entropy-based feature selection with modern gradient-boosting algorithms to have a workable, and interpretable, tool to predict student stress. This study represents a methodological pilot investigating which psychological domains best discriminate self-reported stress levels, rather than a deployable clinical prediction system. The empirical findings suggest that this strategy works: CatBoost reached 97.00% accuracy and 1.000 area-under-the-curve on the held-out test set and cross-validation yielded 92.8% accuracy (±3.31%), providing a more conservative and generalizable performance estimate than the 97% test-set accuracy. One interesting feature of the results is the feature importance analysis. Items that measured motivation and goal orientation always received greater scores of importance in comparison with the items measuring academic workload, therefore, were more powerful predictors of stress category. This observation provides a practical observation that is parallel to the current psychological theory. The next step in the work in the future would be the validation of the framework with the help of the clinical stress assessment tools, expansion of the dataset to students in various geographic locations, and school types, and the investigation of the model stability under the conditions of the long-term conditions, i.e., assessing the possibility of the model trained on students of a specific cohort to be applied to the students of the next year without re-training.

The predictive model should be combined with real-time monitoring systems and the creation of individual intervention pathways in the long-term perspective is logically the continuation of this research.

References

[1] Ovi MS, Hossain J, Rahi MR, Akter F. Protecting Student Mental Health with a Context-Aware Machine Learning Framework for Stress Monitoring. 2025. Arxiv preprint: https://arxiv.org/pdf/2508.01105.

[2] Suraj Arya S, Ramli NA. Predicting the Stress Level of Students Using Supervised Machine Learning and Artificial Neural Network (ANN). Indian J Eng. 2024;21:e9ije1684.

[3] Shahapur SS, Chitti P, Patil S, Nerurkar CA, Shivannagol VS, et al. Decoding Minds: Estimation of Stress Level in Students Using Machine Learning. Indian J Sci Technol. 2024;17:2002-2012.

[4] Khosla A. Multimodal Analysis and Predictive Modeling for Mental Health Detection Through Lifestyle and Behavioral Data Using Machine Learning. J Inf Syst Eng Manag. 2025;10:2046-2063.

[5] Nellore H. Data Analytics and Machine Learning Approaches for Predicting Academic Stress and Enhancing Student Wellbeing. Int j adv res innov. 2025;13:23-29.

[6] Ahuja R, Banga A. Mental Stress Detection in University Students Using Machine Learning Algorithms. Procedia Comput Sci. 2019;152:349-353.

[7] Ding C, Zhang Y, Ding T. A Systematic Hybrid Machine Learning Approach for Stress Prediction. Peer J Comput Sci. 2023;9:e1154.

[8] Singh A, Singh K, Kumar A, Shrivastava A, Kumar S. Machine Learning Algorithms for Detecting Mental Stress in College Students. 2024. arXiv preprint: https://arxiv.org/pdf/2412.07415

[9] Mohamed ES, Naqishbandi TA, Bukhari SA, Rauf I, Sawrikar V, et al. A Hybrid Mental Health Prediction Model Using Support Vector Machine, Multilayer Perceptron, and Random Forest Algorithms. Healthc Anal. 2023;3:100185.

[10] Suryawanshi NS. Predicting Mental Health Outcomes Using Wearable Device Data and Machine Learning. Int J Innov Sci Res Technol. 2021;6:1334-1341.

[11] Manjunath P, Twinkle S, Shreya P, Ashok V, Sultana S. Predictive Analysis of Student Stress Level Using Machine Learning. Int J Eng Res Technol. 2021;9:76-80.

[12] Kumar M, Singh N, Wadhwa J, Singh P, Kumar G, et al. Utilizing Random Forest and Xgboost Data Mining Algorithms for Anticipating Students' Academic Performance. Int J Mod Educ Comput Sci. 2024;16:29-44.

[13] Hasan ME, Arif M, Rakibul Hasan SM, Muwanguzi M, Abaatyo J, et al. Prevalence, Associated Factors, and Machine Learning-Based Prediction of Depression, Anxiety, and Stress Among University Students: A Cross-Sectional Study From Bangladesh. J Health Popul Nutr. 2025;44:1-9.

[14] Kutlimuratov A, Achilov B, Seitnazarov K, Allayarov P, Saymanov I, et al. XGBoost Ensemble Algorithm for Classifying Tomato Leaf Diseases Based on Texture Descriptors. Agri Engineering. 2026;8:98.