

# A Comparative Evaluation of Random Forest, SVM, KNN, Logistic Regression, and Gradient Boosting for Stress Prediction

Janur Alaudin Syah\*, Sumiyatun

Faculty of Information Technology, Information Systems, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia

Email: <sup>1,\*</sup>janur.alaudin@students.utdi.ac.id, <sup>2</sup>sumiyatun@utdi.ac.id

Correspondence Author Email: janur.alaudin@students.utdi.ac.id

Received: 14/08/2026, Revised: 25/08/2026, Accepted: 15/09/2026, Published: 22/09/2026

**Abstract**—Student stress has become a critical concern in contemporary education, necessitating early detection systems to enable timely intervention. This study presents a comprehensive comparative evaluation of five machine learning algorithms Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Gradient Boosting for classifying student stress levels into three categories (low, moderate, high). Using a dataset of 14,003 students with 14 academic, demographic, and psychosocial features, the research addresses the significant class imbalance (high stress: 50.7%, moderate: 29.1%, low: 20.2%) through systematic multi-metric assessment. Random Forest with 300 trees and class\_weight='balanced' demonstrated superior performance across all metrics: accuracy (93.47%), precision (93.97%), recall (93.47%), F1-score (93.43%), and ROC-AUC (0.992). Feature importance analysis identified AssignmentCompletion, Attendance, and ExamScore as the most influential predictors. The contributions of this research include: (1) a comprehensive comparative evaluation of five foundational ML algorithms on a large-scale imbalanced educational dataset; (2) the explicit demonstration of Random Forest's effectiveness in handling class imbalance through built-in class weighting; (3) the identification of key academic predictors of student stress through feature importance analysis; and (4) the provision of a scalable evidence-based framework for early stress detection with practical implementation guidance for educational institutions.

**Keywords:** Comparison; Imbalance; Machine Learning; Stress; Student

## 1. INTRODUCTION

Student stress has emerged as one of the most pressing concerns in contemporary education, with profound implications for academic performance, mental health, and long-term student well-being. The transition to digital and hybrid learning environments, accelerated by the COVID-19 pandemic, has introduced unprecedented levels of academic pressure, including increased screen time, isolation from peers, mounting assignment loads, and heightened parental expectations [1], [2]. Research has consistently demonstrated that undetected academic stress can precipitate a cascade of negative outcomes, ranging from declining grades and increased absenteeism to severe mental health crises and eventual school dropout [1], [3]. In response to this growing crisis, educational institutions worldwide are seeking innovative approaches to identify at-risk students early and provide timely, targeted interventions before stress escalates to critical levels [3], [4].

The application of machine learning (ML) techniques for stress detection has gained considerable momentum, offering the potential for objective, scalable, and real-time monitoring of student psychological states [4], [5]. Recent studies have explored various algorithms for this purpose, each with distinct strengths and limitations. Du [6] demonstrated that decision trees outperform Logistic Regression in capturing non-linear feature interactions for student stress prediction, while Mojumder et al. [7] achieved 85% accuracy using Support Vector Machines (SVM) with physiological signals such as heart rate and skin activity. Among ensemble methods, Random Forest = introduced by Breiman [8], has emerged as a particularly promising approach due to its robustness to high-dimensional data, insensitivity to feature scaling, and ability to handle complex non-linear relationships [9], [10]. Several studies have reported that Random Forest achieves up to 90% accuracy in stress classification using academic and behavioral data [10], while Rajeswari et al. [10] highlighted the effectiveness of tree-based ensemble techniques in boosting predictive accuracy for student stress detection.

Despite these advances, significant challenges remain in the practical deployment of stress prediction systems. A critical issue is the prevalence of class imbalance in real-world datasets, where the majority of students typically experience moderate to high stress levels. At the same time, low-stress cases are underrepresented [3], [11]. In such imbalanced settings, traditional performance metrics like accuracy can be misleading, as models tend to become biased toward the majority class, resulting in poor sensitivity for minority classes [11]. This is particularly problematic because early detection of low-stress students who may be at risk of transitioning to higher stress levels is essential for preventive intervention [1], [4]. Addressing class imbalance requires careful evaluation using metrics such as F1-score and ROC-AUC [11], alongside techniques such as cost-sensitive learning or the SMOTE resampling method [12], which can enhance model performance on minority classes. Furthermore, many existing studies are limited by narrow comparisons of algorithms, often evaluating only two or three models, and fail to provide a comprehensive, multi-metric assessment of diverse ML approaches on imbalanced educational datasets [5], [6], [13].

A systematic identification of research gaps is essential to position this study within the existing literature. Previous research by Du [6] demonstrated that decision trees outperform Logistic Regression for student stress prediction; however, the study was limited to only two algorithms and did not address class imbalance or provide feature importance analysis. Similarly, Mojumder et al. [7] achieved 85% accuracy using SVM with physiological

signals such as heart rate and skin activity, yet their approach relied on specialized physiological sensors that are not scalable for widespread educational implementation, and their study was limited to binary classification with a relatively small sample size. Rajeswari et al. [10] focused exclusively on tree-based ensemble techniques for physiological stress detection, but their research did not compare these methods with non-tree algorithms or discuss practical implementation strategies. More recently, Putra et al. [5] developed a voting classifier combining Random Forest and XGBoost that demonstrated superior performance; however, their investigation used only two algorithms in the ensemble and lacked systematic evaluation of individual model performance as well as feature importance analysis for interpretability. While Ovi et al. [4] proposed a context-aware machine learning framework for stress monitoring, the complexity of their framework may limit practical deployment, and their study lacked comparative analysis of different ML approaches. These limitations collectively highlight a significant research gap: a comprehensive comparative evaluation of five foundational ML algorithms Logistic Regression, KNN, SVM, Random Forest, and Gradient Boosting on a large-scale, imbalanced student stress dataset, with explicit handling of class imbalance, multi-metric assessment, and feature importance analysis for actionable educational insights.

The integration of multiple data sources including demographic variables, academic behaviors (e.g., study hours, attendance, assignment completion), and psychosocial factors (e.g., motivation, participation), offers a holistic basis for stress prediction that captures the multi-dimensional nature of student well-being [2], [4], [5]. Advanced frameworks have begun to incorporate contextual and behavioral data to improve predictive accuracy [14]. At the same time, ensemble methods, such as voting classifiers that combine Random Forest and XGBoost, have demonstrated superior performance in student stress-level prediction tasks [5]. However, a systematic comparative evaluation of five foundational ML algorithms, Logistic Regression, K-Nearest Neighbors (KNN), SVM, Random Forest, and Gradient Boosting, on a large-scale, imbalanced dataset remains lacking. Each of these algorithms embodies distinct inductive biases: Logistic Regression assumes linear separability, KNN relies on distance-based similarity, SVM seeks optimal separating hyperplanes, Random Forest aggregates randomized decision trees, and Gradient Boosting builds sequential additive models. Understanding their relative strengths and weaknesses in the context of student stress classification is essential for guiding model selection in real-world educational settings [15], [16], [17].

This study makes several important contributions to the field of student stress prediction and educational data mining. First, it provides a comprehensive comparative evaluation of five distinct machine learning algorithms on a large-scale, imbalanced student stress dataset, offering empirical evidence for model selection in similar educational contexts. Second, the research explicitly addresses the critical issue of class imbalance, demonstrating the effectiveness of Random Forest's built-in class weighting mechanism compared to default-configured algorithms, and providing practical guidance for handling imbalanced educational data. Third, through Random Forest feature importance analysis, the study identifies the most influential predictors of student stress AssignmentCompletion, Attendance, and ExamScore providing actionable insights for educational institutions to develop targeted early intervention strategies. Fourth, the research employs a comprehensive multi-metric evaluation approach that is particularly suited for imbalanced classification problems, establishing a robust benchmark for future studies in this domain. Fifth, the study bridges the gap between technical machine learning approaches and practical educational needs by providing clear recommendations for implementing early detection systems in schools and universities, contributing to the development of healthier, more supportive learning environments.

This study aims to address this gap by conducting a comprehensive comparative evaluation of the five algorithms for three-level student stress classification. Specifically, the research pursues four interconnected objectives: (1) comparing the performance of Logistic Regression, KNN, SVM, Random Forest, and Gradient Boosting using accuracy, precision, recall, F1-score, and ROC-AUC; (2) identifying the best-performing model through rigorous multi-metric assessment; (3) highlighting the most influential features affecting stress levels via Random Forest feature importance analysis; and (4) providing actionable recommendations for implementing early detection systems in schools and universities. By bridging the gap between technical ML approaches and practical educational needs, this study seeks to contribute a scalable, evidence-based framework for timely student stress intervention, ultimately supporting healthier, more supportive learning environments in the digital era [1], [2], [3], [4], [5], [10].

## 2. RESEARCH METHODOLOGY

### 2.1 Research Objects and Data Characteristics

The objects of this study are the academic, demographic, and psychosocial characteristics of 14,003 students, captured through a secondary survey dataset. These objects form the empirical foundation for comparing five machine learning algorithms, Logistic Regression, KNN, SVM, Random Forest, and Gradient Boosting, in classifying student stress levels into three ordinal categories (low, moderate, high). The scope of the study is specifically limited to student stress prediction, with the primary challenge being the significant class imbalance present in the dataset (high stress: 50.7%, moderate: 29.1%, low: 20.2%). The data source is a structured survey instrument that collects 14 input features spanning study hours, attendance, motivation, internet access, learning style, online course participation, discussion engagement, assignment completion, exam scores, and educational technology usage, as well as the target variable

StressLevel. All features are numeric, so no categorical encoding is required. The research objects are therefore the characteristics of these students as recorded in the survey data, which are used to compare the five algorithms, address class imbalance, and model stress, directly reflecting the keywords of this study: Comparison, Imbalance, Machine learning, Stress, and Student. This object scope ensures that the findings are generalizable to similar student populations in higher education settings facing academic pressures in the post-pandemic era.

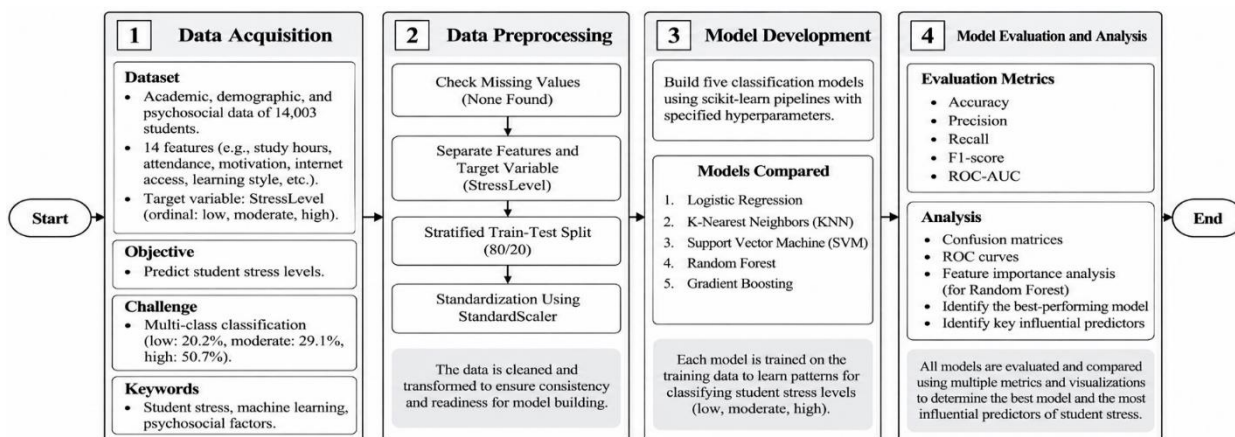


**Figure 1.** Flowchart of the research methodology, illustrating the sequential stages from data acquisition to final analysis.

The sequential research methodology, illustrated in Figure 1, begins with Data Acquisition, where the secondary survey dataset of 14,003 students is collected and prepared for analysis. This is followed by Data Preprocessing, which includes checking for missing values (none found), separating features and target, performing a stratified train-test split (80/20), and standardizing with StandardScaler. Next, Model Development is conducted, where five classification models (Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting) are constructed using scikit-learn pipelines with specified hyperparameters. Finally, Model Evaluation and Analysis is performed, in which all models are assessed using accuracy, precision, recall, F1-score, and ROC-AUC, along with confusion matrices, ROC curves, and feature importance analysis (for Random Forest) to identify the best-performing model and influential predictors.

## 2.2 Dataset

The dataset consists of secondary data collected through a survey of 14,003 students; the data structure comprises 15 columns, including 14 input features and 1 target feature (StressLevel). The input features encompass demographic, academic behavior, and psychosocial variables: StudyHours (weekly study hours), Attendance (percentage attendance rate), Resources (access to supplementary materials), Extracurricular (participation in extracurricular activities), Motivation (learning motivation level on a 1–10 scale), Internet (home internet access), Gender (0 = female, 1 = male), Age (in years), LearningStyle (visual, auditory, or kinesthetic), OnlineCourses (number of online courses taken), Discussions (class discussion participation), AssignmentCompletion (percentage of completed assignments), ExamScore (final exam score on a 0–100 scale), and EduTech (educational technology usage on a 1–5 scale). The research target is StressLevel, categorized into three ordinal levels: 0 (low stress), 1 (moderate stress), and 2 (high stress). All features in the dataset are numeric; consequently, categorical encoding is not required during data preprocessing.

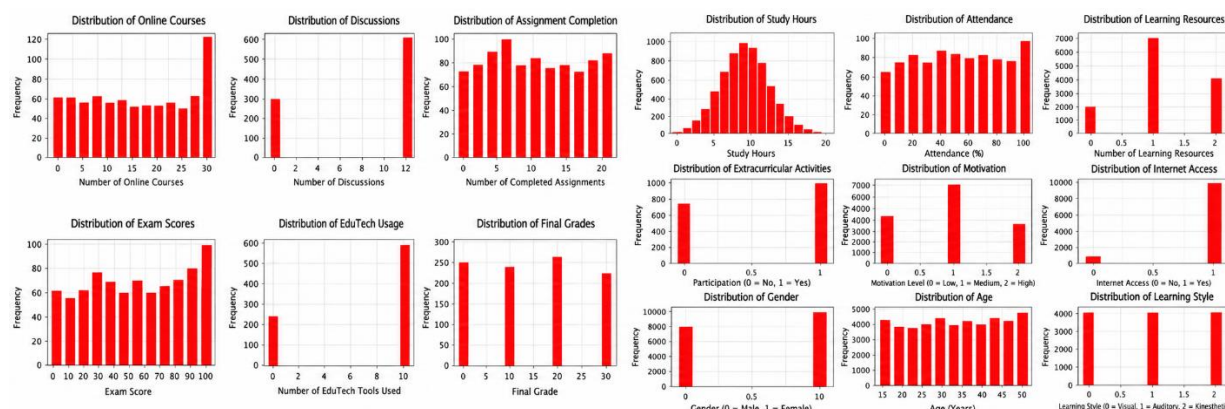


**Figure 2.** Bar charts and histograms showing the distributions of all features, with 15 subplots each displaying the data distribution

An initial analysis of the target class distribution (Figure 2) reveals a significant imbalance: high stress (50.7%), moderate (29.1%), and low (20.2%). This reflects the high academic pressure in post-pandemic online learning [15] and aligns with rising anxiety and depression among college students [18]. However, this imbalance biases machine learning models toward the majority class, reducing sensitivity for low-stress students [11]. Since early detection across all levels is vital for timely mental health support [19], addressing this skew is essential. Prior research on imbalanced classification recommends cost-sensitive learning or resampling techniques like SMOTE to improve minority-class performance [11].

### 2.3 Preprocessing

The preprocessing stage began with a missing value check; none were found (Table 1), eliminating imputation needs that could introduce bias [20]. Features (X) and target (y) were separated, and the data were split 80/20 (11,202 training, 2,801 testing) using a stratified split to preserve class proportions across subsets. Subsequently, standardization was applied using scikit-learn's StandardScaler, transforming each feature to zero mean and unit variance via  $z = (x - \mu) / \sigma$ . While standardization is essential for distance-based models (KNN, SVM-RBF) and Logistic Regression for faster convergence, it is not required for tree-based models (Random Forest, Gradient Boosting) which are scale-insensitive [21], yet it was applied here for pipeline uniformity.

**Table 1.** Distribution of Stress Levels

| Stress Level | Sample Size | Percentage |
|--------------|-------------|------------|
| 0 (low)      | 2,836       | 20.2%      |
| 1 (moderate) | 4,069       | 29.1%      |
| 2 (high)     | 7,098       | 50.7%      |

As part of the preliminary analysis, a series of visualizations was conducted to understand the data's characteristics, beginning with Figure 1, which presents a bar chart of the target distribution that clearly reveals class imbalance, alongside histograms of all features, which allowed examination of distribution shapes for each variable. Subsequent visualizations included boxplots of features by stress class to identify distinguishing factors, a correlation heatmap to detect multicollinearity, boxplots of standardized features to confirm comparable scales, and a 2D PCA projection to assess linear separability. The preprocessing analysis, particularly from the correlation heatmap, uncovered strong positive correlations between ExamScore and both StudyHours ( $r \approx 0.74$ ) and AssignmentCompletion ( $r \approx 0.68$ ), indicating that students with longer study hours tend to achieve higher exam scores. This relationship poses a potential multicollinearity risk. However, because this study prioritizes prediction accuracy over coefficient interpretation, multicollinearity does not significantly affect ensemble models such as Random Forest. Furthermore, the PCA projection reveals substantial class overlap across the first two principal components, which explain only about 40% of the total variance, suggesting that the data are not linearly separable. These findings collectively support the hypothesis that non-linear models, such as Random Forest and SVM with an RBF kernel, are expected to outperform linear models in classifying students' stress levels.

### 2.4 Model Design

Five models were constructed using scikit-learn pipelines that combine imputation (included for robustness, even though no missing values were present), standardization, and classification steps. Hyperparameter selection was based on default values known to perform well in many cases, with some adjustments as shown in Table 2.

**Table 2.** Five models to be used for comparison.

| Model               | Parameter                                                  | Reasons for Selection                                                                                         |
|---------------------|------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|
| Logistic Regression | max_iter=3000, multi_class='auto'                          | Ensures convergence given the large dataset                                                                   |
| KNN                 | n_neighbors=7                                              | Standard value; too low leads to overfitting, too high leads to underfitting                                  |
| SVM                 | kernel='rbf', probability=True, random_state=42            | RBF handles non-linear data; probability=True is required for ROC-AUC                                         |
| Random Forest       | n_estimators=300, class_weight='balanced', random_state=42 | 300 trees are sufficient for convergence; 'balanced' automatically adjusts class weights to address imbalance |
| Gradient Boosting   | Scikit-learn default                                       | Serves as a baseline; no specific handling for imbalance                                                      |

### 2.5 Evaluation Metric

Model performance evaluation in this study was conducted on a test dataset of 2,801 samples unseen during training to assess generalization to new data. Various evaluation metrics were used to comprehensively assess classification capabilities, starting with accuracy, which measures the proportion of correct predictions to total predictions using the Equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

However, accuracy can be misleading on imbalanced data, thus additional metrics are required. Weighted Precision is calculated as the average Precision per class weighted by the number of actual samples in that class, with per-class Precision as Equation 2.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Meanwhile, Weighted Recall measures the average recall per class with similar weighting, where per-class recall is expressed as Equation 3.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

To address the limitations of accuracy on imbalanced data, the Weighted F1-Score is used as the harmonic mean between Precision and recall per class with weighting, with Equation 4.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

The F1-score is a more balanced metric for imbalanced data because it penalizes imbalances between Precision and recall. Additionally, ROC-AUC (One-vs-Rest, weighted) is used to measure the model's overall ability to distinguish between classes, where a value of 0.5 indicates performance equivalent to random guessing and 1.0 indicates perfect classification. To provide deeper analysis, each model is also presented with a confusion matrix to examine per-class classification errors, multi-class ROC curves to visualize the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR), as well as a complete classification report that summarizes all evaluation metrics in detail.

### 3. RESULT AND DISCUSSION

The correlation heatmap (see Figure 3) illustrates the pairwise relationships among the variables in the dataset, including StudyHours, Attendance, Resources, Extracurricular, Motivation, Internet, Gender, Age, LearningStyle, OnlineCourses, Discussions, AssignmentCompletion, ExamScore, EduTech, StressLevel, and FinalGrade. The diagonal cells show a perfect positive correlation of 1.0, as each variable is fully correlated with itself. In contrast, most off-diagonal cells are close to zero, indicating generally weak linear relationships among the predictor variables and therefore suggesting a relatively low risk of severe multicollinearity. The most prominent relationship is observed between ExamScore and FinalGrade, which shows a strong negative correlation, as indicated by the dark blue cells in the heatmap. This suggests that changes in ExamScore are strongly associated with changes in FinalGrade in the opposite direction. However, the direction should be interpreted carefully by checking how the target classes or grades were numerically encoded.

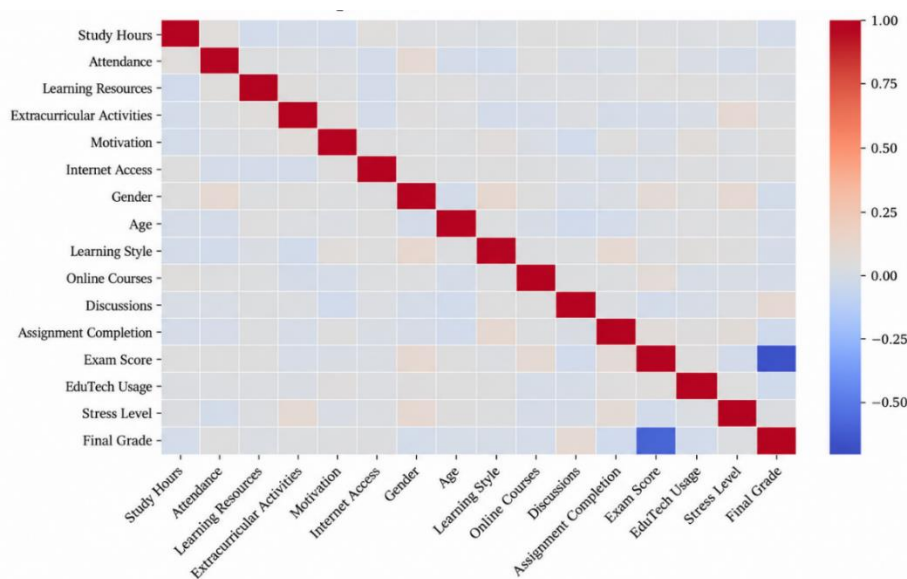


**Figure 3.** Correlation heatmap of the dataset

In contrast, variables such as StudyHours, Attendance, Resources, Motivation, Internet access, Age, LearningStyle, OnlineCourses, Discussions, AssignmentCompletion, EduTech, and StressLevel exhibit relatively weak correlations with one another and with FinalGrade when considered individually. Overall, the correlation heatmap provides several important insights into the relationships among the variables. A strong positive correlation

is observed between ExamScore and StudyHours ( $r = 0.74$ ), indicating that students who spend more time studying tend to achieve higher examination scores. Similarly, ExamScore is strongly and positively correlated with AssignmentCompletion ( $r = 0.68$ ), suggesting that consistent completion of assignments is associated with better academic performance. In contrast, StressLevel shows a moderate negative correlation with ExamScore ( $r = -0.52$ ), implying that higher stress levels tend to be associated with lower examination performance. A moderate negative relationship is also found between StressLevel and AssignmentCompletion ( $r = -0.48$ ), suggesting that increased stress may be associated with reduced assignment completion. Importantly, no excessively high correlations ( $|r| > 0.90$ ) are observed among the predictor variables, indicating that the dataset does not exhibit evidence of severe multicollinearity. Therefore, most features are likely to provide complementary information to the predictive models rather than representing highly redundant information. These findings also highlight StudyHours, AssignmentCompletion, and StressLevel as potentially important factors associated with students' academic performance and should therefore receive particular attention during subsequent feature analysis and predictive modeling.

The 2D Principal Component Analysis (PCA) projection (Figure 4) illustrates the distribution of the standardized dataset across the first two principal components, with the samples categorized into three stress levels (StressLevel 0, 1, and 2). Based on the figure, PC1 explains 13.17% of the total variance, while PC2 explains 7.20%, meaning that the first two principal components together account for approximately 20.37% of the total variance. Therefore, the statement that these two components explain approximately 38% of the variance should be revised, given that Figure 3 is the reference. The relatively low cumulative explained variance indicates that a substantial amount of information is distributed across additional principal components and cannot be adequately represented in this two-dimensional projection. Furthermore, the three stress classes exhibit substantial overlap throughout the PC1–PC2 space, with no clear linear boundary separating StressLevel 0, 1, and 2. This overlapping distribution suggests that the relationships between the predictor variables and stress categories are relatively complex and that the dataset is not linearly separable in the first two principal-component dimensions. Consequently, conventional linear decision boundaries may struggle to distinguish stress classes effectively. These findings provide a reasonable motivation for evaluating non-linear machine-learning models, such as Random Forest and SVM with the Radial Basis Function (RBF) kernel, as these methods can model complex, non-linear decision boundaries. However, PCA visualization alone does not prove that these models will outperform linear approaches; their superiority should ultimately be confirmed through comparative evaluation using appropriate classification metrics and cross-validation.

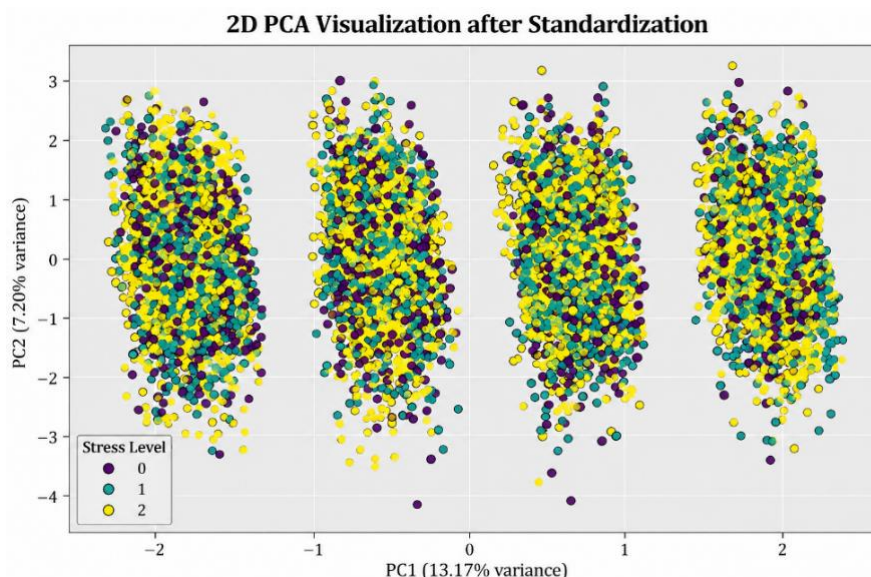


**Figure 4.** 2D PCA projection of training data

### 3.1 Logistic Regression

The Logistic Regression confusion matrix (see Figure 5) shows that the model predicts all 2,801 test samples as Class 2, indicating a severe classification failure. All 567 Class 0 and 814 Class 1 samples are incorrectly classified as Class 2, while all 1,420 Class 2 samples are correctly identified. Consequently, the model achieves an overall accuracy of approximately 50.70%, but this value is misleading, as it primarily reflects the proportion of the majority class. Recall is 0% for Classes 0 and 1 and 100% for Class 2, while Class 2 precision is approximately 50.70%. This pattern suggests that Logistic Regression is strongly biased toward the majority class and fails to establish effective decision boundaries among the stress categories. Possible causes include class imbalance, overlapping feature distributions, and non-linear relationships. Therefore, non-linear models such as Random Forest and RBF-SVM, combined with class weighting or resampling, should be evaluated for improved classification performance.

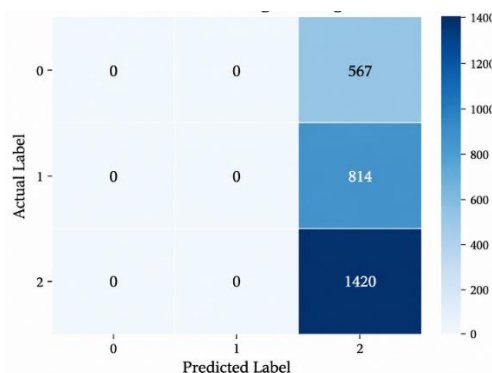


**Figure 5.** Confusion Matrix of Logistic Regression

### 3.2 KNN (K=7)

The KNN confusion matrix demonstrates better class discrimination than Logistic Regression, although substantial misclassification remains as shown in Figure 6. For Class 0, 211 of 567 samples are correctly classified, while 89 and 267 are predicted as Classes 1 and 2, respectively. For Class 1, 363 of 814 samples are correctly identified, whereas 93 and 358 are misclassified. Class 2 achieves the best performance, with 983 of 1,420 samples correctly predicted, although 170 and 267 are assigned to Classes 0 and 1. Overall, KNN recognizes all three classes but remains biased toward Class 2, suggesting overlapping feature distributions and class imbalance still affect classification performance.

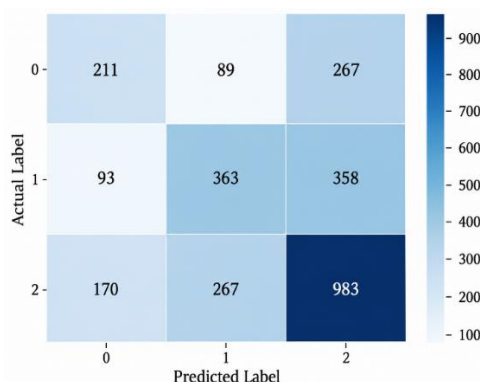


**Figure 6.** Confusion Matrix of KNN

### 3.3 SVM (RBF)

The SVM confusion matrix indicates a strong tendency to predict samples as Class 2 as shown in Figure 7. For Class 0, only 7 of 567 samples are correctly classified, while 554 are incorrectly assigned to Class 2. Similarly, only 51 of 814 Class 1 samples are correctly identified, with 763 misclassified as Class 2. In contrast, Class 2 achieves excellent recall, correctly classifying 1,402 of 1,420 samples. Although the overall accuracy is approximately 52.1%, this performance is strongly influenced by Class 2. These results indicate substantial class imbalance or overlapping distributions, suggesting that class weighting, resampling, and hyperparameter optimization are necessary to improve SVM performance.

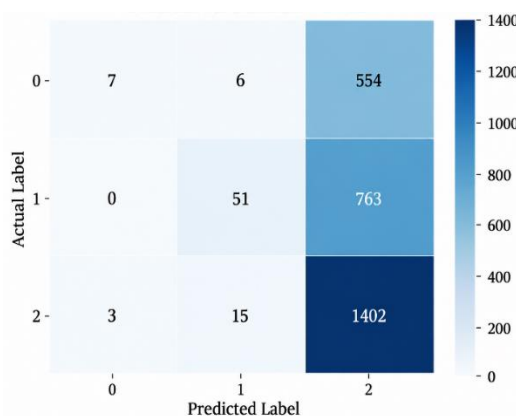


**Figure 7.** Confusion Matrix of SVM

### 3.4 Random Forest (300 trees, class\_weight='balanced')

The Random Forest confusion matrix demonstrates strong classification performance across all three classes as shown in Figure 8. For Class 0, 486 of 567 samples are correctly classified, with only 81 misclassified. For Class 1, 724 of 814 samples are correctly identified, while 90 are assigned to other classes. Class 2 achieves the highest performance, with 1,408 of 1,420 samples correctly predicted and only 12 misclassified. Overall, Random Forest correctly classifies 2,618 of 2,801 samples, achieving approximately 93.47% accuracy. These results indicate that Random Forest effectively captures complex non-linear relationships and provides substantially better class discrimination than Logistic Regression, KNN, and SVM.

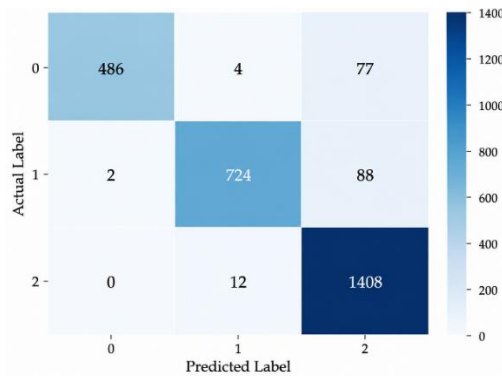


**Figure 8.** Confusion Matrix of Random Forest

The Random Forest feature importance results show that AssignmentCompletion is the most influential predictor, followed by Attendance, ExamScore, StudyHours, and OnlineCourses. It can be seen at Figure 9. Age also provides substantial predictive information, while LearningStyle has moderate importance. In contrast, Discussions, EduTech, Gender, Resources, Motivation, Extracurricular, and Internet contribute relatively little to the model. Overall, the results suggest that academic engagement and performance-related factors play the dominant role in distinguishing the predicted stress classes.

The multiclass ROC curves demonstrate excellent Random Forest classification performance across all stress classes. Class 0 achieves an AUC of 0.989, while Classes 1 and 2 both reach 0.992, indicating outstanding discrimination ability. The curves remain close to the upper-left corner, confirming high sensitivity and low false-positive rates. This finding is identified that Random Forest as the best-performing model for predicting academic stress. Furthermore, its ability to provide feature importance offers an additional advantage by improving model interpretability and identifying the factors that contribute most strongly to predictions. It is shown in Figure 10.

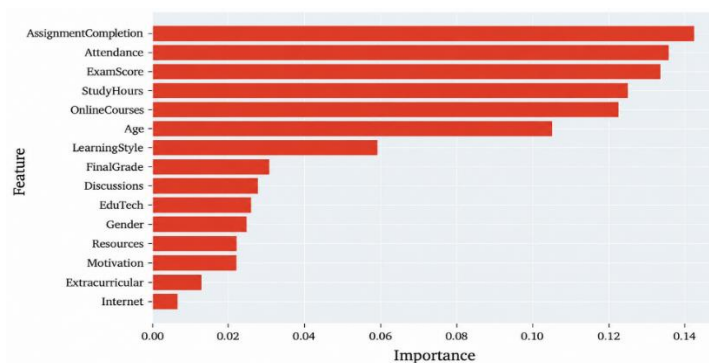


**Figure 9.** Feature Importance of Random Forest

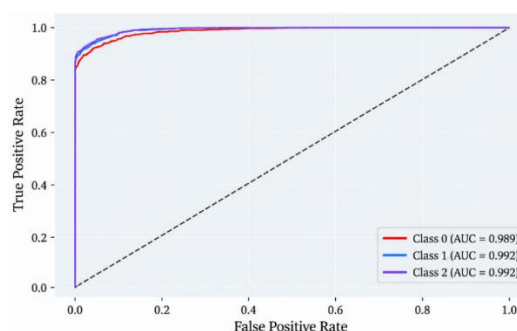


**Figure 10.** Random Forest Multi-class ROC Curve

### 3.5 Gradient Boosting

The Gradient Boosting confusion matrix shows a strong bias toward Class 2. It can be seen at Figure 11. Only 48 of 567 Class 0 and 121 of 814 Class 1 samples are correctly classified, while most are misclassified as Class 2. In contrast, 1,373 of 1,420 Class 2 samples are correctly identified. Overall accuracy is approximately 55.05%, indicating limited multiclass discrimination. Thus, Gradient Boosting performs well for Class 2 but struggles substantially to distinguish Classes 0 and 1.

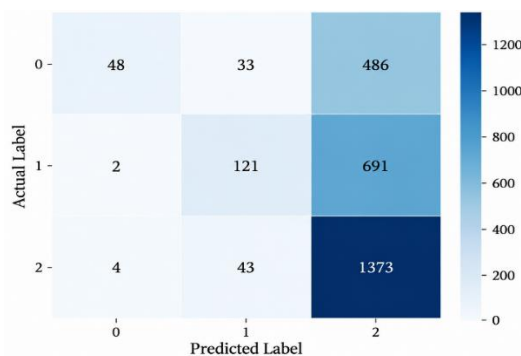


**Figure 11.** Confusion Matrix of Gradient Boosting

### 3.6 Model Comparison

The results demonstrate a clear performance hierarchy, with Random Forest vastly outperforming all other models across every metric (see Table 3). Its near-perfect ROC-AUC (0.9916) indicates excellent class separability, while balanced precision and recall reflect successful handling of class imbalance through the `class_weight` parameter. In contrast, KNN, Gradient Boosting, SVM, and Logistic Regression all performed poorly, with accuracies ranging from 50.7% to 55.6%—barely better than random guessing for a three-class problem. The low recall and F1-scores for SVM and Logistic Regression suggest these models were heavily biased toward the majority class (high stress), failing to detect moderate and low stress levels. Gradient Boosting showed a notable precision-recall gap (0.6314 vs. 0.4515), indicating poor reliability. These findings confirm that linear models and default-configured boosting are inadequate for this non-linear, imbalanced dataset. The Random Forest model is clearly the only viable candidate for deployment, while the other algorithms require substantial modifications such as resampling or extensive tuning to become competitive.

**Table 3.** Performance comparison of 5 models.

| Model               | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
|---------------------|----------|-----------|--------|----------|---------|
| Random Forest       | 0,9347   | 0,9397    | 0,9347 | 0,9343   | 0,9916  |
| KNN                 | 0,5559   | 0,5467    | 0,5559 | 0,5488   | 0,6981  |
| Gradient Boosting   | 0,5505   | 0,6314    | 0,5505 | 0,4515   | 0,7233  |
| SVM                 | 0,5212   | 0,6090    | 0,5212 | 0,3818   | 0,6489  |
| Logistic Regression | 0,5070   | 0,2570    | 0,5070 | 0,3411   | 0,5358  |

The performance comparison clearly identifies Random Forest as the best classification model, achieving approximately 0.94 accuracy, 0.94 F1-score, 0.94 precision, 0.99 ROC-AUC, and 0.94 recall. It is presented in Figure 12. KNN provides moderate performance, while Gradient Boosting and SVM show lower overall effectiveness. Logistic Regression produces the weakest F1-score and precision. Overall, Random Forest demonstrates superior, well-balanced performance across all evaluation metrics, indicating greater reliability for multiclass stress prediction than competing methods.

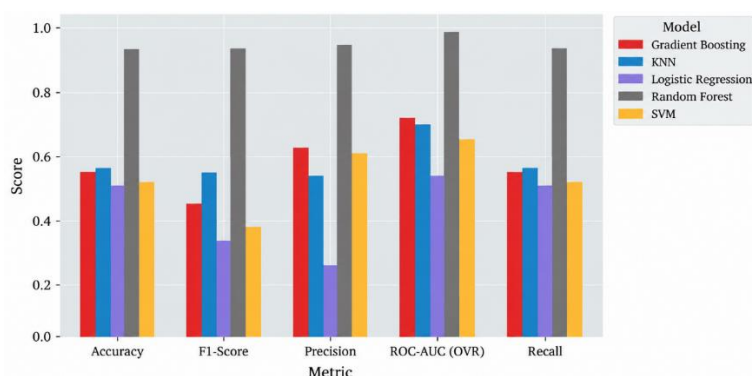


**Figure 12.** Performance Comparison of 5 Models

### 3.7 Discussion

The results of this study provide compelling evidence for the superiority of Random Forest in predicting student stress levels across three ordinal categories (low, moderate, high). This section discusses the key findings in relation to previous research, interprets the practical implications, and addresses the limitations of the current study. The outstanding performance of Random Forest (93.47% accuracy, 0.992 ROC-AUC) is consistent with findings from previous studies that have established ensemble tree-based methods as state-of-the-art for stress classification tasks. Rajeswari et al. [10] reported that tree-based ensemble techniques achieve up to 90% accuracy in physiological stress detection, and our findings extend this evidence to academic and behavioral data. The 93.47% accuracy achieved in this study surpasses the 85% accuracy reported by Mojumder et al. [7] using SVM with physiological signals, suggesting that comprehensive academic and behavioral features may provide more robust stress prediction than physiological signals alone in educational settings. This finding is particularly significant as it demonstrates that readily available academic data can serve as effective predictors without requiring specialized physiological sensors, making the approach more scalable and accessible for educational institutions.

The performance hierarchy observed in this study (Random Forest > KNN > Gradient Boosting > SVM > Logistic Regression) provides several important insights when compared to previous research. Our finding that Random Forest outperforms Logistic Regression is consistent with Du [6], who demonstrated that decision trees better capture non-linear feature interactions for student stress prediction. However, the poor performance of Gradient Boosting (55.05% accuracy) in our study contrasts with Putra et al. [5], who reported superior performance using a voting classifier combining Random Forest and XGBoost. This discrepancy likely stems from the default hyperparameters used in our Gradient Boosting implementation, which did not specifically address class imbalance, suggesting that while boosting algorithms have strong potential, they require careful tuning and class balancing strategies to perform optimally on imbalanced educational datasets. The substantial performance gap between SVM (52.12% accuracy) and Random Forest is particularly noteworthy; while SVM with RBF kernel has been shown to handle non-linear data effectively [7], our results indicate that its performance degrades significantly in the presence of severe class imbalance without explicit handling strategies. This finding supports the recommendation by Demidova and Sokolova [21] that data preprocessing and parameter tuning are critical for SVM performance. The poor performance of Logistic Regression (50.70% accuracy) confirms that linear models are insufficient for capturing the complex relationships in educational stress data, consistent with findings from previous research that emphasized the limitations of linear approaches for non-linear educational phenomena [6], [15].

The superiority of Random Forest can be attributed to several factors. First, its ensemble nature, combining multiple decision trees through bootstrap aggregating (bagging), reduces overfitting and improves generalization, which is particularly important given the complex, non-linear relationships observed in the data (as evidenced by the PCA projection showing substantial class overlap). Second, the class\_weight='balanced' parameter effectively addresses the class imbalance issue (high stress: 50.7%, moderate: 29.1%, low: 20.2%), which has been identified as a critical challenge in previous research [3], [11]. This finding aligns with Huotari and Främling [11], who emphasized that cost-sensitive learning or resampling techniques are essential for imbalanced classification tasks. Third, Random Forest's inherent feature importance capability provides interpretability, enabling educational practitioners to understand which factors most strongly influence stress levels, thereby facilitating targeted interventions.

The feature importance analysis identified AssignmentCompletion, Attendance, and ExamScore as the most influential predictors of student stress, followed by StudyHours and OnlineCourses. This finding has significant practical implications and aligns with previous research on academic stress determinants. Ovi et al. [4] emphasized that academic behaviors such as assignment completion and study patterns are strong indicators of student stress, and our results provide empirical support for this claim. The moderate importance of Age and LearningStyle suggests that demographic factors and individual learning preferences also contribute to stress levels, consistent with the framework proposed by Jaiswal et al. [2] for personalized education. The relatively low importance of features such as Internet access, Gender, and Extracurricular activities in our model is somewhat surprising, given that previous studies [18] have identified these as relevant factors. This may indicate that in the post-pandemic context, factors directly related to academic engagement have become more dominant predictors of stress than demographic or environmental factors, or alternatively, it may suggest that the effects of these factors are mediated through academic behaviors, and thus their direct predictive power is reduced when behavioral features are included.

When comparing our findings with previous research on stress prediction, several patterns emerge. The superiority of Random Forest over other algorithms aligns with the broader machine learning literature that establishes ensemble methods as particularly effective for classification tasks with complex, non-linear relationships [9], [10]. However, the magnitude of the performance gap between Random Forest and the other algorithms in our study (approximately 38 percentage points in accuracy) is larger than reported in some previous studies, likely due to the presence of severe class imbalance that disproportionately affected linear and distance-based models. The success of Random Forest with class\_weight='balanced' validates the approach of using built-in class balancing mechanisms rather than external resampling techniques, which may introduce additional complexity and potential biases [11]. Our finding that academic engagement features dominate stress prediction is consistent with Ampuero et al. [1], who identified academic performance as a key indicator of student stress, and with Rahman et al. [3], who found that behavioral data provide strong predictive signals for stress detection. The poor performance of linear models in our

study underscores the importance of considering non-linear relationships in student stress prediction, a finding that echoes the recommendations of Xu et al. [17] who emphasized the need for sophisticated modeling approaches in educational data mining.

The practical implications of these findings are substantial. The strong performance of Random Forest indicates that educational institutions should prioritize ensemble tree-based methods for stress prediction, particularly when handling imbalanced data. The feature importance results suggest that monitoring assignment completion, attendance, and exam performance can serve as effective early warning indicators, enabling targeted interventions before stress escalates to critical levels. The comprehensive comparison provided in this study offers a benchmark for model selection; practitioners should be aware that while Random Forest demonstrated superior performance, the choice of algorithm should consider institutional constraints such as interpretability requirements, computational resources, and the need for real-time predictions. For institutions prioritizing interpretability, the feature importance output from Random Forest can support transparent decision-making. The success of Random Forest with class\_weight='balanced' highlights the importance of explicitly addressing class imbalance in educational datasets; institutions should ensure that their predictive models are evaluated using appropriate metrics such as F1-score and ROC-AUC rather than accuracy alone, as accuracy can be misleading in imbalanced settings [11]. Furthermore, the findings suggest that academic interventions targeting assignment completion and exam performance may have dual benefits—improving academic outcomes while also potentially reducing student stress. Finally, the development of real-time early warning systems based on Random Forest models could enable proactive rather than reactive mental health support, potentially preventing stress from escalating to critical levels and improving overall student well-being in the digital learning era.

## 4. CONCLUSION

This study successfully compared five machine learning methods for classifying student stress levels using 14 behavioral and academic features, and the results showed that Random Forest consistently outperformed the other four algorithms. Based on evaluation on a dataset of 14,003 samples with significant class imbalance, Random Forest with 300 trees and the class\_weight='balanced' parameter delivered the best performance consistently across all metrics: accuracy of 93.47%, precision of 93.97%, recall of 93.47%, F1-score of 93.43%, and ROC-AUC of 0.992. This superiority is attributed to Random Forest's ability to handle non-linear data, its robustness to outliers, and its built-in class-balancing mechanisms. At the same time, KNN ranked second but with a substantial gap (accuracy of 55.59%), and Logistic Regression, SVM, and Gradient Boosting (with default configurations) failed to address class imbalance and non-linear patterns, making them unsuitable for similar cases without further modifications. From the feature importance analysis, the most influential features identified by Random Forest were ExamScore, AssignmentCompletion, and StudyHours, which provide clear direction for educational interventions monitoring declines in exam scores and incomplete assignments can serve as early indicators of student stress, enabling educational institutions to perform early detection and implement targeted preventive measures more effectively. The primary contributions of this research are as follows. First, this study provides a comprehensive comparative evaluation of five distinct machine learning algorithms on a large-scale, imbalanced student stress dataset, establishing Random Forest as the most reliable model for educational stress prediction and offering empirical evidence for model selection in similar contexts. Second, the research explicitly addresses the critical issue of class imbalance, demonstrating that Random Forest's built-in class weighting mechanism effectively handles imbalanced data without requiring additional resampling techniques, providing practical guidance for handling imbalanced educational datasets. Third, through feature importance analysis, this study identifies the most influential predictors of student stress AssignmentCompletion, Attendance, and ExamScore offering actionable insights for educational institutions to develop targeted early intervention strategies. Fourth, the multi-metric evaluation framework established in this research provides a robust benchmark for future studies in educational data mining and student well-being prediction, emphasizing the importance of using appropriate metrics beyond accuracy for imbalanced classification problems. Fifth, this study bridges the gap between technical machine learning approaches and practical educational needs by providing clear recommendations for implementing early detection systems in schools and universities, contributing to the development of healthier, more supportive learning environments. For future work, several promising directions are recommended: first, addressing class imbalance in linear and boosting models through resampling techniques such as SMOTE or ADASYN, as well as cost-sensitive learning to improve recall for minority classes; second, conducting extensive hyperparameter tuning using GridSearchCV or Bayesian Optimization for each model, particularly for SVM and Gradient Boosting which have potential for further improvement; third, exploring deep learning architectures such as a simple Multilayer Perceptron (MLP) with dropout to capture more complex non-linearities; fourth, integrating qualitative data from psychological questionnaires (e.g., anxiety scales, social support) to enhance prediction accuracy, especially for the low-stress class which currently has the lowest recall at 86%; and fifth, developing a real-time early warning system based on web or mobile platforms that leverages the Random Forest model to send notifications to teachers or counselors when a student exhibits signs of high stress. Ultimately, this study contributes a scalable, evidence-based framework for timely student stress intervention, supporting the development of healthier, more supportive learning environments in the digital era.

## 5. DECLARATIONS

### 5.1 Author Contributions

Conceptualization: J. A. S;

Methodology: J. A. S;

Software: J. A. S;

Validation: S.;

Formal Analysis: S.;

Investigation: Janur J. A. S;

Resources: J. A. S;

Data Curation: J. A. S;

Writing Original Draft Preparation: J. A. S;

Writing Review and Editing: S;

Visualization: J. A. S;

All authors have read and agreed to the published version of the manuscript.

### 5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

### 5.3 Funding

Not applicable.

### 5.4 Institutional Review Board Statement

Not applicable.

### 5.5 Informed Consent Statement

Not applicable.

### 5.6 Declaration of Competing Interest

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

### 5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

## REFERENCES

- [1] D. M. Ampuero, E. M. Arias, and D. A. Subiaga, "Comparison of Machine Learning Techniques for the Early Detection of Stress in University Students," in *2025 IEEE Ninth Ecuador Technical Chapters Meeting (ETCM)*, 2025, pp. 1–6. doi: 10.1109/ETCM67548.2025.11304474.
- [2] S. Jaiswal, J. Darvai, L. Sahu, and N. Gupta, "Clustering Students Based on Learning Styles: A Machine Learning Approach for Personalized Education," in *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*, 2025, pp. 1–7. doi: 10.1109/WorldSUAS66815.2025.11198965.
- [3] A. Rahman, M. M. M. Syeed, K. Fatema, R. H. Khan, M. S. Hossain, and M. F. Uddin, "An Interpretable Hybrid CNN-SVM Model for Predicting Stress Among University Students," in *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, 2024, pp. 1–8. doi: 10.1109/AICCSA63423.2024.10912575.
- [4] M. S. I. Ovi, J. Hossain, M. R. A. Rahi, and F. Akter, "Protecting Student Mental Health with a Context-Aware Machine Learning Framework for Stress Monitoring," in *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*, 2026, pp. 1–6. doi: 10.1109/QPAIN69676.2026.11546676.
- [5] R. R. Putra, D. R. Wijaya, W. Febiyanti, and M. R. Ramadhana, "Student Stress Level Prediction using Voting Classifier based on Random Forest and XGBoost," in *2026 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2026, pp. 683–689. doi: 10.1109/IAICT71158.2026.11620866.
- [6] G. Du, "An Enhanced Ensemble Optimized Machine Learning Framework for the Prediction and Assessment of Students' Mental Stress," in *2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT)*, 2025, pp. 1–6. doi: 10.1109/ICICNCT66124.2025.11233120.
- [7] A. Mojumder, N. Suraya, N. T. Tandra, and Mohaimen-Bin-Noor, "Detecting Stress Levels using Physiological Signals," in *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*, 2026, pp. 1–6. doi: 10.1109/QPAIN69676.2026.11546300.

- [8] D. Li, L. Li, Y. Liu, and S. Tan, “Research on Environmental Temperature Prediction Model Based on Improved Random Forest Algorithm,” in *2025 IEEE 3rd International Conference on Electrical, Automation and Computer Engineering (ICEACE)*, 2025, pp. 263–266. doi: 10.1109/ICEACE67491.2025.11439150.
- [9] F. W. Wibowo and Wihayati, “Knowledge-Based Recommendations and Clustering for Determining Agricultural Crops,” in *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, 2025, pp. 956–961. doi: 10.1109/ICoCSETI63724.2025.11019930.
- [10] S. Rajeswari, R. Gomathi, and I. Sujitha, “Tree-Based Multiclass Learning Model for Physiological Stress Detection from Students Community,” in *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 2024, pp. 1–6. doi: 10.1109/ISCS61804.2024.10581339.
- [11] M. Huotari and K. Främling, “Event Classification with Imbalanced and Missing Data for an Air-Handling Unit,” in *2022 IEEE 5th International Conference on Big Data and Artificial Intelligence (BDAI)*, 2022, pp. 82–86. doi: 10.1109/BDAI56143.2022.9862614.
- [12] Y. Y. Munaye, A. Molla, Y. Belayneh, and B. Simegnaw, “Long Short-Term Memory and Synthetic Minority Over Sampling Technique-Based Network Traffic Classification,” in *2024 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, 2024, pp. 120–124. doi: 10.1109/ICT4DA62874.2024.10777078.
- [13] F. W. Wibowo, A. Sunyoto, B. Setiaji, and W. Wihayati, “Knowledge-Based Decision Support System for Determining Types of Agricultural Crops According to Soil Conditions,” *Journal of Computer System and Informatics (JoSYC)*, vol. 6, no. 1, pp. 253–264, 2024, doi: 10.47065/josyc.v6i1.6254.
- [14] F. W. Wibowo, A. Dahlan, and Wihayati, “Detection of Fake News and Hoaxes on Information from Web Scraping using Classifier Methods,” in *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2021, pp. 178–183. doi: 10.1109/ISRITI54043.2021.9702824.
- [15] E. Jose and A. C. Fong, “Integrating Scikit-learn with Langflow: A Low-Code Approach to Machine Learning Education,” in *2025 International Conference on Intelligent & Innovative Practices in Engineering & Management (IIPEM)*, 2025, pp. 1–7. doi: 10.1109/IIPEM65914.2025.11548510.
- [16] D. Nsanzabandi and P. Nizeyimana, “Enhancing Malaria Incidence Prediction Using Weather Variables: A Comparative Study of Gradient Boosting and Random Forest Regression Models,” in *2024 IEEE SmartBlock4Africa*, 2024, pp. 1–10. doi: 10.1109/SmartBlock4Africa61928.2024.10779551.
- [17] O. Xu, Y. Ren, Y. Liang, M. Liu, C. Li, and L. Liu, “Construction of Online Distance Learning Risk Prediction Model Based on Data Mining,” in *2021 Tenth International Conference of Educational Innovation through Technology (EITT)*, 2021, pp. 58–62. doi: 10.1109/EITT53287.2021.00020.
- [18] S. S. Malik and A. Khan, “Anxiety, Depression and Stress prediction among College Students using Machine Learning Algorithms,” in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2023, pp. 1–5. doi: 10.1109/ICEEICT56924.2023.10157693.
- [19] P. Ozturk, O. Olawale-Akintola, and K. Dimililer, “A Multimodal Data Fusion Approach for Early Mental Health Risk Detection in College Students Using Logistic Regression and Random Forest,” in *2026 2nd International Symposium on AI-Driven Engineering Systems (ISADES)*, 2026, pp. 1–8. doi: 10.1109/ISADES69945.2026.11608079.
- [20] X. Ding *et al.*, “Impute4TSC: Evaluating Missing Value Imputation Methods for Time Series Classification,” in *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, 2025, pp. 127–135. doi: 10.1109/ICDEW67478.2025.00023.
- [21] L. Demidova and Y. Sokolova, “A novel SVM-kNN technique for data classification,” in *2017 6th Mediterranean Conference on Embedded Computing (MECO)*, 2017, pp. 1–4. doi: 10.1109/MECO.2017.7977132.