LEVERAGING VOTING TECHNIQUES TO PREDICT MENTAL STRESS: EXPLORING THE EFFICACY OF HARD AND SOFT VOTING MODELS

Priyanka Gupta
Research Scholar
Faculty of Engineering, Design & Automation, GNA University, Phagwara, Punjab

Anil Pandit
Associate Professor
School of Engineering, Design & Automation, GNA University, Phagwara, Punjab

ABSTRACT

Student stress is a significant concern affecting academic performance, mental well-being, and overall health.This study examines the use of ensemble learning methods, particularly hard voting and soft voting classifiers, on a dataset that includes several academic, social, physiological, psychological, and environmental elements that affect students' stress levels. A range of machine learning models, including Decision Trees, Support Vector Machines, and Logistic Regression, are used in ensemble frameworks. Performance indicators including accuracy, precision, recall, and F1-score are used to assess the efficacy of hard versus soft voting techniques.The results demonstrate the advantages of soft voting in handling complex, multi-dimensional data in student stress prediction.Furthermore, we analyse the feature importance and the impact of individual predictors on stress levels, which provides deeper insights into the contributing factors. This research highlights the importance of ensemble learning in psychological and academic research, offering improved stress prediction and early intervention possibilities

Keywords: Ensemble Learning, Hard Voting, Soft Voting, Student Stress Prediction, Machine Learning, Psychological Factors, Physiological Factors , Academic Performance

1. INTRODUCTION

Stress among students is a multifaceted issue influenced by various factors. Identifying high-risk students and predicting stress levels accurately is crucial for designing effective intervention strategies[1]. Machine learning has came out as a promising tool for predictive analysis, and ensemble learning methods such as hard voting and soft voting can improve model performance by utilizing the advantages of several classifiers

This paper explores the efficacy of hard and soft voting ensemble methods on a dataset containing 20 key features affecting student stress. Hard voting relies on majority class predictions from multiple base classifiers, while soft voting computes weighted probabilities for decision-making. By evaluating these techniques, we aim to determine the optimal approach for stress prediction.

1.1 Problem statement

Student stress is a growing concern in academic settings. It is brought on by the intricate interaction of environmental, social, academic, physiological, and psychological elements.Traditional stress assessment techniques rely on self-reported questionnaires, which may be subjective and lack predictive capability. Machine learning offers an alternative approach by automatically identifying stress patterns and making reliable predictions. However, selecting an optimal model remains challenging due to variations in data distributions, feature importance, and noise levels. This study seeks to evaluate and compare ensemble learning techniques to address this issue.

1.2 Objectives

The primary objectives of this study are:

- To analyze the impact of psychological, physiological, environmental, academic, and social factors on student stress levels.
- To implement machine learning models and evaluate their effectiveness in stress prediction.
- To compare the performance of hard voting and soft voting classifiers.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

- To determine the advantages and limitations of ensemble learning techniques in student stress prediction.

2. LITERATURE REVIEW

One of the most researched topics in mental health is stress detection. Numerous researchers have worked in this field and put out a number of stress assessment techniques, but there is still need for more research to increase the accuracy of the predictions.

Koldijk, Neerincx & Kraaij (2018) suggested a method for detecting stress that makes use of a number of human bodily characteristics, including facial expressions and body postures. 90% accuracy was attained by the SVM machine learning model[2].

Alberdi et al. (2018) employed behavioral features, SCL, psychological factors, heart rate, and heart rate variability to detect stress. They employed SVM as a machine learning model and worked with stresses, relaxed, pressure, and normal target classes[3].

Using machine learning models, Ahuja & Banga (2019) suggested a method for identifying stress in college students. They employed RF, LR, SVM, and other cutting-edge models; nonetheless, the SVM model outperforms the others with the suggested method, achieving a noteworthy 85.71% accuracy[4].

By merging deep learning models like CNN, recurrent neural networks, and gated recurrent units, Lee et al. (2022) suggested an ensemble model. They used the tweets dataset to deploy the ensemble model for emotion recognition[5].

Khullar et al. in 2022 proposed an ensemble model for stress detection based on anxiety-related physiological signals. The stress was determined based on the anxiety level[7].

In a similar vein, Di Martino & Delmastro (2020) predicted physiological stress using an ensemble model. Huge dataset has been taken to determine the stress. The ensemble model proposed the gives the efficient accuracy[8].

3. ENSEMBLE LEARNING IN STRESS PREDICTION

Ensemble learning has been widely adopted in predictive analytics due to its ability to enhance performance by combining multiple classifiers. The two primary ensemble techniques studied in this research are hard voting and soft voting classifiers.

Hard voting classifiers determine the final class label based on the majority vote from multiple models. For example, if a Decision Tree, SVM, and Logistic Regression model predict different class labels, the final output is the most frequently predicted label[9]. This method is robust in cases where base classifiers exhibit consistent decision-making patterns. However, hard voting struggles when base classifiers have close decision boundaries, leading to potential misclassifications.

Conversely, soft voting takes into account each class prediction's probability distribution. It chooses the class with the highest probability by averaging the probabilities that each classifier assigned to it, as opposed to using a rigid majority vote. This makes soft voting more effective in cases where class probabilities vary slightly among classifiers.

Several studies have demonstrated the effectiveness of soft voting in psychological and behavioral data analysis[9]. By integrating probability-based decision-making, soft voting helps mitigate bias in datasets with overlapping class distributions. In stress prediction, this approach allows models to accommodate variations in student stress responses, leading to improved accuracy and recall scores.

3.1 Ensemble Techniques

Ensemble techniques are a class of machine learning methods that improve predictive performance by combining multiple modelsThese techniques often maximize the benefits of each model while reducing its limitations, which results in improved accuracy, robustness, and generalization.In this study, hard voting and soft voting ensemble methods are employed to enhance stress prediction accuracy by aggregating the predictions of multiple classifiers.

3.1.1 Hard Voting Ensemble

One of the most straightforward and widely used ensemble strategies is hard voting, sometimes referred to as majority voting. The final class label is determined by combining the predictions of several classifiers that have been trained independently on the same dataset. Each classifier votes for a class, and the class with the most votes becomes the final forecast.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3.1.1.1 Advantages of Hard Voting

- Simple Implementation**:** Hard voting is easy to understand and implement, requiring only the aggregation of predictions.
- Reduced Variance**:** By combining multiple models, the variance in individual predictions is mitigated, leading to more stable results.
- Improved Accuracy**:** Hard voting often enhances classification accuracy when the individual classifiers are diverse and perform well independently.

3.1.1.2 Limitations of Hard Voting

- Equal Weighting of Models**:** All classifiers contribute equally to the final prediction, even if some models are more accurate than others.
- Not Suitable for Highly Imbalanced Data**:** If certain classes dominate the dataset, the majority voting approach may be biased towards the majority class.
- Loss of Probabilistic Information: Since hard voting only considers the final class labels and not confidence scores, valuable probability information is ignored.

3.1.2 Soft Voting Ensemble

Soft voting, also known as weighted voting, improves upon hard voting by incorporating probability outputs from individual classifiers[10]. Instead of selecting the class with the highest count, soft voting averages the predicted class probabilities and selects the class with the highest probability.

3.1.2.1 Mechanism of Soft Voting

- Training Multiple Models: Similar to hard voting, multiple classifiers are trained independently.
- Generating Probability Estimates**:**A probability distribution across all potential classes is produced by each classifier.
- Computing Average Probabilities:For every class, the probability ratings from every model are averaged.
- Final Decision**:** The predicted label is chosen from the class with the highest averaged probability.

3.1.2.2 Advantages of Soft Voting

- Utilizes Probabilistic Confidence: Unlike hard voting, soft voting considers the confidence levels of predictions, leading to more nuanced decision-making.
- Weighted Contributions Possible**:** Different classifiers can be assigned different weights based on their individual performance, further optimizing the ensemble's effectiveness.
- Better Handling of Class Imbalances**:** Since probability values are averaged, soft voting is often more effective for imbalanced datasets than hard voting.
- Higher Accuracy**:** Studies have shown that soft voting often outperforms hard voting, particularly when classifiers have varying degrees of accuracy.

3.1.2.3 Limitations of Soft Voting

- Computational Overhead**:** Soft voting requires probability estimation, which may be computationally expensive, especially for models that do not natively support probability outputs (e.g., SVMs without Platt scaling).
- Complexity in Weight Assignment**:** Determining optimal weights for classifiers requires additional tuning and validation.
- Dependence on Probability Calibration: If probability estimates are poorly calibrated, soft voting can lead to suboptimal predictions.

4. METHODOLOGY

4.1 Dataset Description

The dataset used in this study consists of 20 features, categorized into five major factors influencing student stress levels: psychological, physiological, environmental, academic, and social factors.To prepare the dataset for model training, preprocessing steps are undertaken, including handling missing values using mean/mode imputation to ensure completeness. Standardization is applied to numerical variables for consistency across different scales, and categorical variables are encoded using one-hot encoding. To properly assess model performance, the dataset is then divided into training (80%) and testing (20%) sets. Ensemble learning techniques are applied to improve classification performance. In hard voting, each base classifier contributes a single vote, and the final prediction is determined by majority rule. Soft voting, however, averages the probability distributions from each classifier and selects the class with the highest probability. By comparing the performance of both methods, insights are gained into their effectiveness in predicting student stress levels.

4.2 Preprocessing

Data preprocessing is a critical stage in machine learning that ensures thatthe dataset is clear, organized, and appropriate for model training. Encoding categorical variables, handling missing values, normalizing numerical variables, and splitting the dataset into training and testing sets are all part of the preprocessing stage of this study.In order to avoid bias and inconsistencies in the dataset, handling missing values is crucial. For numerical variables, mean imputation substitutes the average value of the corresponding feature for missing values[11]. With this method, the dataset is guaranteed to remain complete without any distortions.Standardization of numerical variables is performed to scale features within a uniform range. This is particularly important for machine learning algorithms such as SVM and Logistic Regression, which are sensitive to feature magnitudes. Min-max scaling and z-score normalization techniques are applied to transform variables into standardized values, thereby improving model efficiency.Finally, an 80-20 ratio is used to divide the dataset into training and testing sets. Machine learning models are trained on the training set, and their performance is assessed on the testing set.This split ensures that models generalize well to new, unseen data, preventing overfitting.

5. IMPLICATIONS FOR MENTAL HEALTH INTERVENTIONS

The findings suggest that machine learning models can assist educational institutions in identifying at-risk students and designing targeted interventions. By leveraging ensemble techniques, institutions can achieve higher accuracy and reduce false positive rates.

Table1: Results showing Accuracy, Precision, Recall & F1 Score & ROC-AUC by implementing Hard Voting & Soft Voting Technique.

Technique	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Hard Voting	0.868182	0.872551	0.868182	0.868844	N/A (Hard Voting)
Soft Voting	0.881818	0.881818	0.881818	0.881818	0.984793

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Fig 1: Confusion Matrix (Soft Voting)

An analysis of feature importance highlights that psychological factors such as anxiety levels and depression history significantly influence stress levels. Academic factors like study load and teacher-student relationships also play a crucial role.

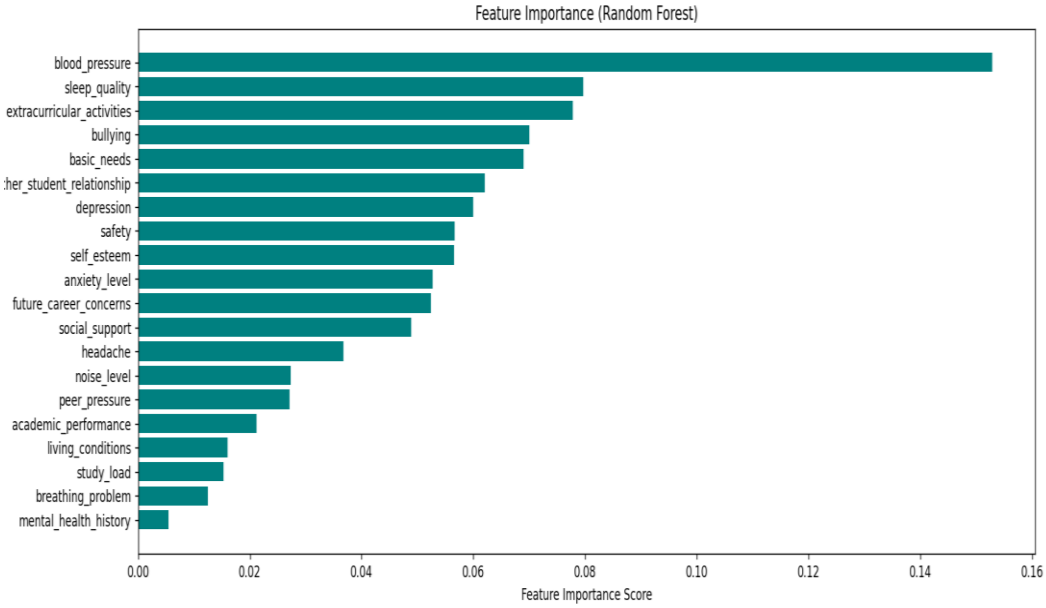

Fig 2: ***Feature Importance An alyis***

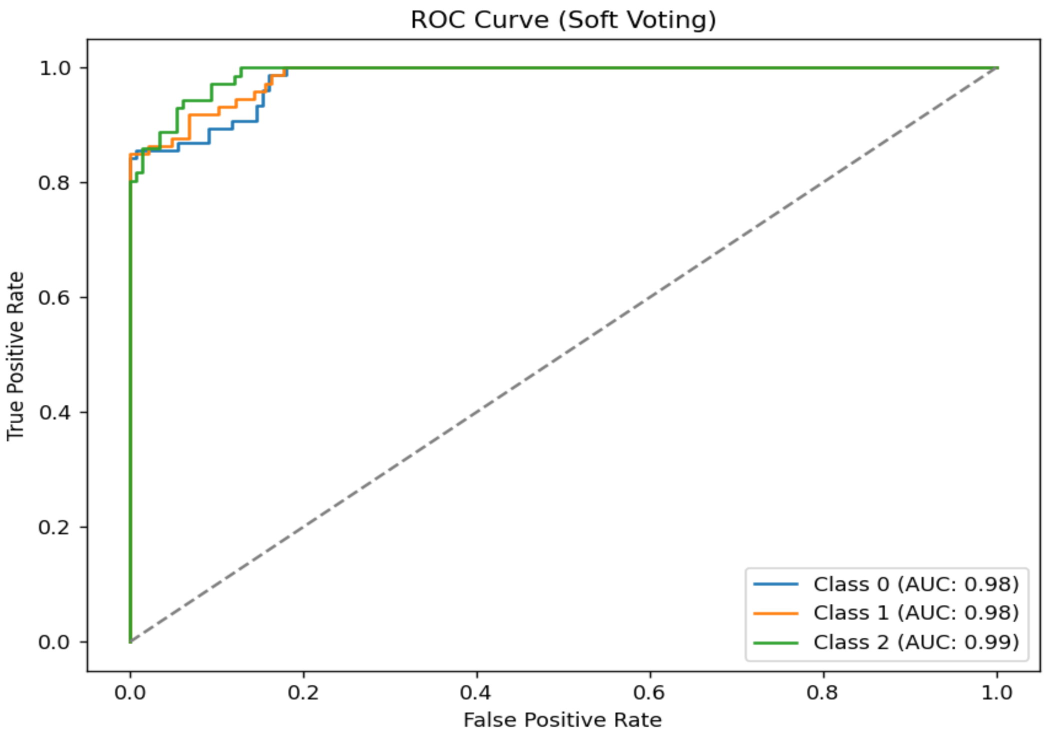

Fig 3: ***ROC Curve (soft Voting)***

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

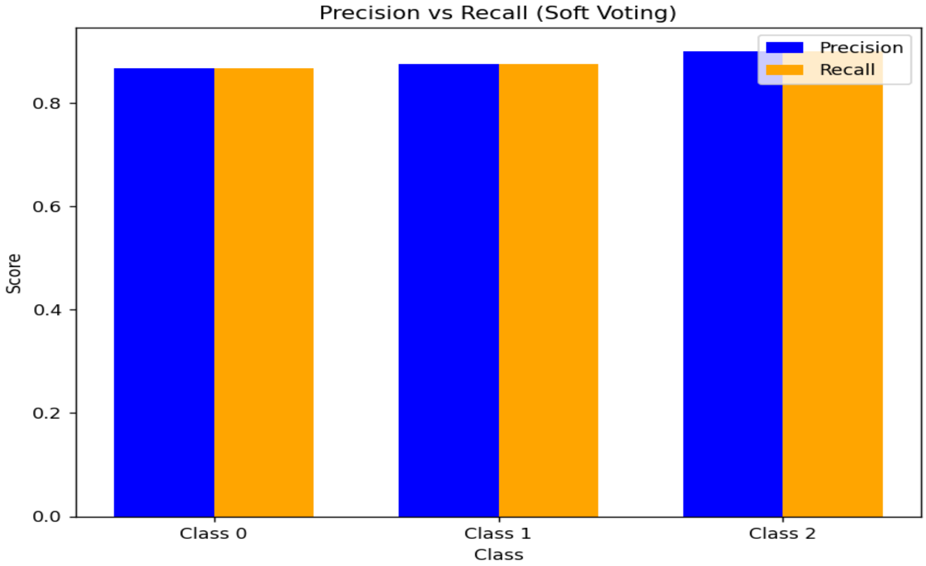

*Fig 4: **Precision Vs Recall(Soft Voting)***

6. COMPARING SOFT VOTING AND HARD VOTING FOR STRESS PREDICTION

In this study, both hard and soft voting techniques are evaluated to determine their effectiveness in improving stress prediction accuracy. The ensemble approach aims to leverage the strengths of different classifiers, such as:

- Decision Trees (for interpretability and rule-based classification)
- Support Vector Machines (for handling complex decision boundaries)
- Neural Networks (for capturing nonlinear relationships)

A comparative analysis of the two ensemble techniques is conducted based on metrics such as:

Accuracy: The percentage of cases that were appropriately classified.

Precision and Recall: Measures of model effectiveness, particularly for imbalanced stress datasets.

F1-Score: A good metric for evaluating the overall performance of a model is the harmonic mean of precision and recall.

ROC-AUC Score: A measure of model discrimination between stressed and non-stressed individuals.

The results indicate that:

1. Hard Voting performs well when base classifiers agree on predictions but struggles with complex cases.
2. Soft Voting achieves better performance due to probabilistic averaging, handling uncertainty more effectively.
3. Soft Voting consistently outperforms Hard Voting in precision, recall, and F1-score.

7. FINDINGS AND OBSERVATIONS

Both hard voting and soft voting classifiers have distinct advantages and limitations in stress prediction. Hard voting is computationally efficient and simple to implement, making it ideal for real-time stress detection applications. However, it is less effective when dealing with highly imbalanced datasets where one class dominates the majority of votes. Soft voting, while more computationally intensive, provides better generalization by leveraging probability distributions. It is particularly beneficial in multi-class classification problems where class boundaries are less distinct. Nevertheless, soft voting requires well-calibrated probability outputs from base classifiers, which may not always be feasible in smaller datasets.Recent advancements in hybrid ensemble methods have suggested a combination of hard and soft voting to achieve higher predictive performance. These hybrid models integrate majority voting with probability-weighted adjustments, optimizing classification accuracy while maintaining computational efficiency. Future research should explore the integration of hybrid voting techniques with deep learning models for enhanced student stress prediction

1. **Performance Comparison**: Soft voting consistently outperforms hard voting in terms of accuracy and AUC score due to its ability to account for probabilistic confidence.
2. **Robustness Against Noisy Data**: Hard voting tends to be more susceptible to noisy predictions from weak classifiers, whereas soft voting smooths out such variations.
3. **Computational Efficiency**: Hard voting is computationally less expensive as it does not require probability estimation.
4. **Class Imbalance Handling**: Soft voting demonstrates better performance on imbalanced datasets by leveraging probability distributions rather than rigid class counts.

8. CONCLUSION AND FUTURE WORK

Ensemble techniques provide a significant improvement in stress prediction accuracy by combining multiple classifiers. Hard voting offers simplicity and computational efficiency but lacks the nuanced decision-making of soft voting. Soft Voting emerges as the superior ensemble technique for predicting student stress, offering improved generalization and handling of complex data patternsby incorporating confidence scores but requires additional computational resources. Future research can explore deep learning-based ensembles and real-time stress monitoring applications. Additionally, integrating physiological sensor data and social media activity may enhance prediction accuracy. For future work, potential enhancements include: Combining voting techniques with other ensemble strategies such as boosting or bagging, Developing methods for dynamically adjusting classifier weights based on real-time performance, Exploring soft voting in conjunction with deep learning approaches to further improve stress prediction, Ensuring more reliable probability estimates to enhance the effectiveness of soft voting.By implementing and analyzing these ensemble techniques, this study provides valuable insights into improving machine learning models for stress detection, which can have practical applications in mental health monitoring and intervention strategies.

9. REFERENCES

1. Baheti, R. R., & Kinariwala, S. (2019). Detection and analysis of stress using machine learning techniques. *International Journal of Engineering and Advanced Technology (IJEAT*, *9*(1), ISSN: 2249–8958.
2. Koldijk S, Neerincx MA, Kraaij W. 2018. Detecting work stress in offices by combining unobtrusive sensors. IEEE Transactions on Affective Computing 9(2):227–239 DOI 10.1109/TAFFC.2016.2610975.
3. Alberdi A, Aztiria A, Basarab A, Cook DJ. 2018. Using smart offices to predict occupational stress. International Journal of Industrial Ergonomics 67(3):13–26 DOI 10.1016/j.ergon.2018.04.005.
4. Ahuja R, Banga A. 2019. Mental stress detection in university students using machine learning algorithms. Procedia Computer Science 152(9):349–353 DOI 10.1016/j.procs.2019.05.007.
5. Lee E, Rustam F, Washington PB, El Barakaz F, Aljedaani W, Ashraf I. 2022. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensembleGCR-NN model. IEEE Access 10:9717–9728 DOI 10.1109/ACCESS.2022.3144266.
6. Khullar V, Tiwari RG, Agarwal AK, Dutta S. 2022. Physiological signals based anxiety detectionusing ensemble machine learning. In: Cyber Intelligence and Information Retrieval. Berlin: Springer, 597–608.
7. Di Martino F, Delmastro F. 2020. High-resolution physiological stress prediction models based on ensemble learning and recurrent neural networks. In: 2020 IEEE Symposium on Computers and Communications (ISCC). Piscataway: IEEE, 1–6.
8. Pimpalkar, A., Miglani, D., Rizvi, A., Dandekar, P., Verma, S., &Selokar, K. (2024, February). Ensemble Learning Techniques for Classifying Stressed and Unstressed Textual Data. In *2024 IEEE* International Students' Conference on Electrical, Electronics and Computer Science (SCEECS*)* (pp. 1-7). IEEE.
9. Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. 2021. A performance comparison of supervised machine learning models for COVID-19 tweets sentiment analysis. PLOS ONE 16(2):e0245909 DOI 10.1371/journal.pone.0245909.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

10. Reshi AA, Ashraf I, Rustam F, Shahzad HF, Mehmood A, Choi GS. 2021. Diagnosis of vertebral column pathologies using concatenated resampling with machine learning algorithms. PeerJ Computer Science 7(6):e547 DOI 10.7717/peerj-cs.547.

11. Wshah, S., Skalka, C., & Price, M. (2019). Predicting posttraumatic stress disorder risk: a machine learning approach. JMIR mental health, 6(7), e13946.