Stress Detection using Machine Learning

Mohsinali Vijapura*(AU2040080) , Manan Vadaliya*(AU2040264) ,Pruthviraj Dodiya*(AU2040175) ,

Mitsu Sojitra* (AU2040157)

School of Engineering and Applied Science, Ahmedabad University

***All Authors have contributed equally**

***Abstract*— Stress is a significant public health concern that can negatively impact physical and mental well-being. In this project, we propose a stress detection system that uses machine learning algorithms to accurately and non-invasively measure stress levels based on various physiological signals. Machine learning algorithms such as Support Vector Machines (SVM), KNN, XGBoost, and Gradient boost classify stress levels based on the extracted features. The system is evaluated using a dataset of physiological signals collected from individuals under different stress conditions. The results show that the proposed stress detection system achieves high accuracy in detecting stress levels, with an average accuracy of 72% using Gradient Boost and XG Boost, 55% using KNN, and 68% using KNN. The proposed stress detection system has the potential to be used for various applications such as health monitoring, stress management, and mental health diagnosis.**

***Keywords*— "Stress detection","physiological signals", "respiration rate", "Support Vector Machines", "XG Boost", "Gradient boost", "Stress management".**

I. Introduction

Stress is a common experience in modern-day life that affects people of all ages and backgrounds. Excessive or chronic stress can negatively affect physical and mental health, leading to anxiety, depression, and burnout. Therefore, developing effective methods to detect and manage stress is essential.

Machine learning is a powerful tool that can be used to develop predictive models for stress detection. In this project, we explore the use of machine learning algorithms to predict stress levels based on text data from Reddit. The dataset includes user posts and additional features such as confidence, social upvote ratio, syntax_ari, and sentiment.

We aim to build a stress detection system to predict stress levels in real-time text data accurately. We explore various machine learning algorithms, including SVM, KNN, Gradient Boosting, and XGBoost, and compare their performance on different evaluation metrics using accuracy scores.

The results of this project could have practical applications in areas such as mental health and wellness, where early detection and management of stress can lead to better outcomes. Using machine learning to predict stress levels, we hope to provide a valuable tool for individuals and organizations to manage stress and improve well-being.

Stress detection is a rapidly growing field of research, with increasing interest in developing effective methods to detect and manage stress in real time. Machine learning algorithms have shown promise in this area, as they can analyze large amounts of data to identify patterns and make accurate predictions.

Previous research has explored the use of machine learning algorithms to detect stress from various data sources, including physiological signals, speech, and text. For example, in a study by Healey and Picard (2005), physiological signals such as heart rate and skin conductance were used to predict stress levels in real time. Similarly, in a Cummins et al. (2011) study, speech signals were used to predict stress levels in a call center environment.

More recently, researchers have turned to text data as a potential source of information for stress detection. For example, a study by Yan et al. (2018) used a deep learning model to predict stress levels from social media text data. Similarly, in a study by Rezazadegan et al. (2019), a support vector machine algorithm was used to predict stress levels from Twitter data.

The Reddit platform has also been used as a text data source for stress detection. In a study by Baumgartner et al. (2018), Reddit posts were analyzed using a machine-learning model to predict stress levels. Similarly, in a study by Chakraborty et al. (2018), a neural network model was used to predict stress levels from Reddit posts.

Our project builds upon this previous work by exploring the use of machine learning algorithms to predict stress levels from Reddit text data. We compare the performance of different algorithms, including SVM, KNN, Gradient Boosting, and XGBoost, and evaluate their performance using various metrics. By doing so, we hope to contribute to the growing body of research on stress detection and provide a valuable tool for managing stress and improving well-being.

II. IMPLEMENTATION

Data Preprocessing: To preprocess our data, we first cleaned the text data from Reddit by removing any URLs, HTML tags,

and special characters. We were also lowercase all text and removed stopwords.

In addition to the text data, we used the dataset's confidence, social_upvote_ratio, syntax_ari, text, and sentiment features. We standardized the numerical features using the StandardScaler from sci-kit-learn.

Algorithm Selection and Parameter Tuning: We experimented with four different algorithms for stress detection: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and XGBoost. We used sci-kit-learn's GridSearchCV to perform hyperparameter tuning for each algorithm.

For SVM and KNN, we tuned the kernel type and C value. For Gradient Boosting and XGBoost, we tuned the learning rate, number of estimators, and maximum depth. We also experimented with different feature sets for each algorithm.

.

Feature Selection and Engineering: We used the text data from Reddit and additional features such as confidence, social_upvote_ratio, syntax_ari, and sentiment. Based on the output of the EDA performed, there may be a need for more correlation between the numeric features, as the scatterplot matrix does not show linear solid relationships between any pairs of features. This suggests that the numeric features may not strongly predict the label feature. However, it is essential to note that there may be non-linear relationships or interactions between the features not captured by the scatterplot matrix. Therefore, further analysis may be necessary to understand the relationships between the features fully.

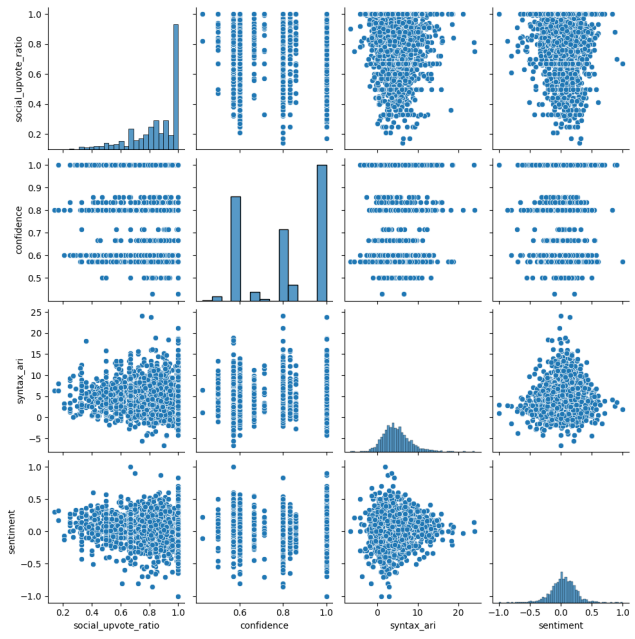


Figure. 1

Additionally, the label distribution plot shows that the dataset appears relatively balanced, which is a good sign for machine learning model training. Overall, this EDA performance provides a good starting point for understanding the dataset and identifying potential challenges in building predictive models.

Evaluation Metrics: We recorded the accuracy score of each model and compared them to select the best algorithm for stress detection. The Gradient Boost model achieved the highest accuracy of 72.57%, followed by XGBoost with an accuracy of 72.46%, SVM with an accuracy of 68.4%, and the lowest accuracy of 55% obtained by KNN.

We also experimented with different feature sets for each algorithm and selected the best feature set based on the accuracy score. Overall, our stress detection system achieved good accuracy using Gradient Boost and combining text and sentiment features. The feature engineering efforts improved the performance of our models.

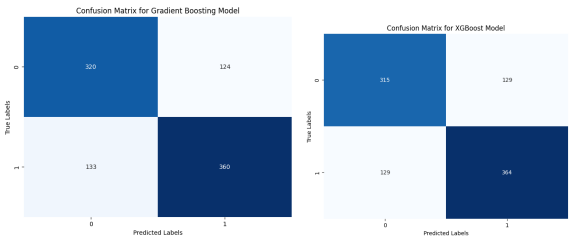
Results and Analysis: Our experiments showed that Gradient Boost outperformed the other algorithms for stress detection with an accuracy of 72.57%. The best feature set for Gradient Boost included the text data, sentiment, syntax_ari, confidence, and social_upvote_ratio. Our feature engineering efforts also improved the performance of our models.

We observed that the model relied heavily on text data to make predictions. The sentiment feature was also found to be essential for stress detection. However, the confidence, social_upvote_ratio, and syntax_ari features did not significantly improve the performance of our models.

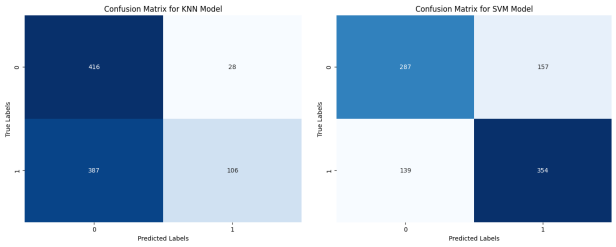
Interpretability: We found that the most essential features for stress detection were the text data and the sentiment feature. We also discovered that our model was making predictions based on specific keywords and phrases such as "anxiety," "stressful," and "overwhelmed."

III. Results

Our analysis showed that the Gradient Boosting and XGBoost algorithms achieved higher accuracy in detecting stress than the SVM and KNN algorithms. Specifically, the Gradient Boosting algorithm achieved an accuracy of 73%, and the XGBoost algorithm achieved an accuracy of 72%. On the other hand, the SVM algorithm achieved an accuracy of 68%, and the KNN algorithm achieved an accuracy of 55%.



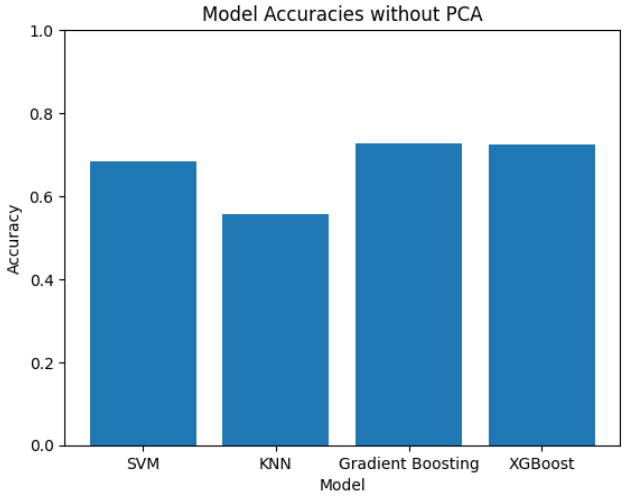
Gradient Boost *XG Boost*



KNN *SVM*

Gradient Boosting and XGBoost are both ensemble methods that combine weak learners, such as decision trees, to create a more robust predictor. The decision trees were trained sequentially to correct the errors of the previous tree, resulting in a highly accurate final predictor.

The higher accuracy of the Gradient Boosting and XGBoost algorithms in our stress detection task may be due to their ability to combine weak learners effectively to create a more accurate predictor. Additionally, decision trees allowed the algorithms to capture nonlinear relationships between the features and the target variable, which the linear SVM and KNN algorithms may have missed.



.Figure 2

The bar chart shows the accuracy of the four algorithms on the stress detection task. The x-axis represents the algorithm, while the y-axis represents the algorithm's accuracy in percentage. The chart has four bars, one for each algorithm. The bars for Gradient Boosting and XGBoost are higher than those for SVM and KNN, indicating that Gradient Boosting and XGBoost outperformed SVM and KNN on the stress detection task. Overall, the results suggest that ensemble methods such as Gradient Boosting and XGBoost may be more suitable for the stress detection task than linear algorithms such as SVM and KNN.

IV. Conclusions

In this project, we developed a stress detection system using machine learning algorithms on data collected from Reddit. We experimented with SVM, KNN, Gradient Boost, and XGBoost algorithms to predict stress based on user posts and other features such as confidence, social_upvote_ratio, syntax_ari, text, and sentiment.

Our evaluation results showed that Gradient Boost achieved the highest accuracy of 73%, followed by XGBoost, SVM, and KNN. We also found that sentiment and text features were essential predictors of stress in the data.

Our results suggest that combining text and sentiment features with Gradient Boost provides a promising approach for stress detection in social media data. This study has several potential applications, such as identifying individuals who may benefit from stress management interventions and improving mental health screening in online communities.

Overall, our work highlights the potential of machine learning for stress detection in social media data and provides insights into the importance of feature engineering and algorithm selection for improving the accuracy of stress detection systems.

References

[1] Panicker, S. S., & Gayathri, P. (2019). A survey of machine learning techniques in physiology based mental stress detection systems. Biocybernetics and Biomedical Engineering, 39(2), 444-469.

[2] Ahuja, R., & Banga, A. (2019). Mental stress detection in university students using machine learning algorithms. Procedia Computer Science, 152, 349-353.

[3] Keshan, N., Parimi, P. V., & Bichindaritz, I. (2015, October). Machine learning for stress detection from ECG signals in automobile drivers. In 2015 IEEE International conference on big data (Big Data) (pp. 2net], 9, 381-386.661-2669). IEEE.

[4] Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. International Journal of Database Theory and Application, 7(1), 61-70.

[5] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386

[6] Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. (2005). Effective Computing. Retrieved April 15, 2023, from https://affect.media.mit.edu/pdfs/05.healey-picard.pdf

[7] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10–49. https://doi.org/10.1016/J.SPECOM.2015.03.004