A Comparison of Word Embedding Techniques for Stress Identification using Machine Learning

K.S.Kalaivani
Department of Artificial Intelligence
Kongu Engineering College
Erode, India
kalaiprani123@gmail.com

C.S.Kanimozhi Selvi
Department of Artificial Intelligence
Kongu Engineering College
Erode, India
kanimozhi@kongu.ac.in

N.Sandeep
Department of Artificial Intelligence
Kongu Engineering College
Erode, India
sandeepn.22@kongu.edu

T.Vasanthasri
Department of Artificial Intelligence
Kongu Engineering College
Erode, India
vasanthasrit.22aid@kongu.edu

S. VarshiniShilin
Department of Artificial Intelligence
Kongu Engineering College
Erode, India
varshinishilins.22@kongu.edu

***Abstract*— In stress analysis, the focus is on examining individuals' language or voice for signs of stress. This study utilizes a dataset obtained from DravidanLangTech-2024 and employs various embedding techniques like Term Frequency-Inverse Document Frequency (TF-IDF), Doc2Vec and Word2Vec. These techniques transform data points into numerical representations in a high-dimensional space, aiding algorithms in understanding relationships and similarities between them. Machine learning algorithms, including Random Forest, Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and Logistic Regression, are crucial in stress analysis. In the Tamil dataset, Random Forest achieved impressive accuracies of 98%, 97%, and 99% for TF-IDF, Doc2Vec, and Word2Vec embeddings, respectively. Similarly, in the Telugu dataset, Random Forest consistently displayed the highest accuracies across all three embeddings. Particularly noteworthy is the exceptional performance of the Random Forest Algorithm with Word2Vec embedding, achieving an outstanding 99% accuracy in predicting stress for both Tamil and Telugu datasets.**

Keywords—Stress Identification, TF_IDF, Doc2Vec, Word2Vec, Machine Learning

I. INTRODUCTION

In the contemporary digital age, social media has emerged as a prominent platform for individuals to express their emotions, thoughts, and experiences. Among the myriad of languages spoken and written on these platforms, Tamil and Telugu stand out as significant contributors, reflecting the rich linguistic diversity of the Indian subcontinent. However, the multilingual nature of social media posts, where individuals frequently blend languages, poses challenges for sentiment analysis, particularly in accurately deciphering emotional nuances.

Our research delves into the intricate realm of sentiment analysis and stress detection, with a specialized focus on multilingual datasets in Tamil and Telugu. These datasets, sourced from social media platforms, present a unique challenge due to the amalgamation of languages, making it more complex to discern emotions accurately.

To address this challenge head-on, we employ a range of machine learning techniques specifically tailored to navigate the complexities of mixed languages and adeptly categorize emotions. Our methodology aims to develop a robust system capable of real-time stress identification, especially in contexts where conventional sentiment analysis methodologies falter.

The significance of our research extends beyond academic curiosity; it holds profound implications for mental health support. Early detection of stress can be a pivotal factor in providing timely intervention and support, potentially mitigating more severe mental health issues down the line. By rigorously testing various machine learning algorithms on multilingual social media content, we aspire to identify optimal strategies for recognizing stress patterns effectively. The objective of our study is as follows:

- Evaluating the effectiveness of different embedding techniques like TF-IDF, Doc2Vec and Word2Vec.
- Assessing the performance of a range of machine learning algorithms including Random Forest, SVM, Decision Tree, Naïve Bayes and Logistic Regression for detecting stress.

The rest of the paper is structured as follows. Section II provides a detailed discussion of related work. Section III and IV covers the proposed methodology and the results & analysis of the trials respectively. Section V presents the conclusion.

II. RELATED WORK

Stress intensified by the COVID-19 pandemic, poses significant health risks and challenges relationships. Leveraging the Stress Analysis on Social Media dataset from Kaggle, the study achieved an 83.74% f1-score with a novel model architecture combining Transformer Encoder layers and Bi-directional-LSTM. An explainable AI approach was also implemented for result interpretation [1]. Today's IT

professionals frequently face stress due to evolving lifestyles and workplace cultures, despite company efforts to improve work environments. This study aims to analyze stress patterns and identify influential variables using machine learning techniques. Data from medical experts' mental health surveys was collected and preprocessed. Various machine learning algorithms, including decision tree, random forest, KNN, logistic regression, and Naive Bayes, were applied to forecast stress threats. The results showcased improved system performance in predicting stress-related issues [2]. The pandemic increased stress levels among university students due to health concerns, increased screen time, and lack of social activities. Using data from 444 students, the study found that 24.1% had high psychological stress and 11.26% had high social stress. Machine learning models achieved an accuracy of 80.5% in predicting stress levels. The best-performing model used the Multilayer Perceptron algorithm with feature reduction and hyperparameter optimization techniques. This study had limitations like self-reported data and small sample size [3].

A questionnaire-based dataset using the 5-point Likert Scale and the Perceived Stress Scale (PSS) was employed. Regression analysis showed Linear Regression (LR) as the best model with an accuracy of 97.8%, while the Logistic Regression Classifier (LRC) achieved 96% accuracy in classification. Despite the study's small dataset and limited attribute mapping, the findings suggest machine learning can effectively automate stress analysis among students during crises [4]. The study tested 16 Hz binaural beats (BBs) to reduce stress. Participants took tests and listened to the beats. The beats improved focus and detection accuracy by 27.08%. The best machine learning model, SVM had 82.5% accuracy in classifying mental states. However, self-reported stress levels didn't always match these results. EEG showed brain activity changes during tasks and recovery. Future research could explore different beat frequencies and use deep learning for better results.[5] . There are 6023 journal entries from 755 Q-Life app users to understand resilience factors. Using sentiment analysis, Support Vector Machine (SVM) showed the best performance with 89.7% accuracy in classifying positive or negative sentiments. Thematic analysis revealed 14 negative themes like stress and 13 positive themes like gratitude, along with 7 coping mechanisms including time management and mindfulness [6].

College students face various mental health issues like stress, anxiety, and depression, impacting both their personal growth and campus stability. There's a need for a comprehensive system to predict and manage student stress. Factors contributing to stress include workload, assignments, and personal issues. Machine learning can leverage past data to accurately predict stress levels and offer personalized suggestions to manage it [7]. The study focuses on using machine learning to identify and classify mental health conditions like depression, stress, and anxiety in children and adolescents during the Covid-19 pandemic. Using the DASS-21 Scale and data from 2050 Chinese participants, the research categorized these conditions into Normal, Moderate, and Severe levels. Support Vector Machine (SVM) proved to be the most effective classification algorithm for identifying these mental health conditions. [8]. The study aims to delve into the effectiveness of ensemble methods within supervised learning. By combining multiple algorithms, we seek to optimize classification accuracy and address limitations, especially in managing large datasets. We will explore how leveraging the strengths of different algorithms can compensate for individual weaknesses, ultimately aiming to improve overall performance in supervised learning tasks.[9]

Utilizing a dataset of Jordanian students, we aim to predict and classify mental stress levels during the COVID-19 pandemic using machine learning algorithms such as Linear Regression and Logistic Regression. Our focus is on refining prediction accuracy and exploring the impact of isolation and quarantine on stress levels [10]. Further exploration in [11] involves employing additional classifiers such as Random Forest, Gradient Boosting and Gaussian Naive Bayes to enhance stress detection accuracy. Additionally, implementing 10-fold cross-validation will ensure robust model evaluation. The authors in [12] address the challenge of identifying stress in sedentary workers through a multimodal AI framework, achieving 96.09% accuracy in stress detection. Proposed work includes leveraging the SWELL-KW dataset for stress prediction using various fusion algorithms and emphasizing early fusion for stress classification. Additionally, it highlights the importance of monitoring stress levels and task load over time for proactive intervention.

The objective of this research is to create an intelligent system utilizing machine learning algorithms such as Linear Regression, K-Nearest Neighbor, and Support Vector Machine to detect stress based on physiological indicators. The dataset comprises information from 300 individuals aged between 18 to 25, encompassing variables such as heart rate, blood pressure, skin response and gender. By employing Gridsearch for optimization on Google Colab, Support Vector Machine emerged as the optimal classifier, achieving an accuracy score ranging from 95.00% to 96.67% in both training and testing phases [13]. This study [14] focuses on stress detection using physiological parameters like Electrocardiogram, Galvanic Skin Response, etc. Machine learning algorithms including Decision Tree, Naïve Bayes and K-Nearest Neighbor are employed. Results exhibit 100% accuracy across algorithms, indicating successful stress prediction. Graphs and a GUI aid in visualization and interpretation of stress detection outcomes. The authors in [15] address stress in IT professionals using machine learning on OSMI mental health survey 2017 data. Various techniques like Boosting and Decision Trees were applied. Boosting showed the highest accuracy. Prominent stress influencers identified include gender, family history, and workplace health benefits. These findings enable targeted stress reduction strategies in workplaces.

Most of the existing works have compared various machine learning and deep learning models for stress identification. The state-of-the-art works did not attempt to compare different word embedding techniques for this task. This study aims to evaluate the effectiveness of different embedding techniques and various machine learning algorithms such as Random Forest, SVM, Decision Tree, Naïve Bayes and Logistic Regression.

III. PROPOSED WORK

By combining advanced embedding techniques with powerful machine learning algorithms, this study seeks to improve our understanding of the text and make accurate predictions based on it. As shown in figure 1, initially, we transformed words into numerical representations using techniques such as TF-IDF, Word2Vec, and Doc2Vec. These methods help us capture the meaning and context of words within the dataset. Subsequently, we applied different machine learning algorithms like Random Forest, Decision Tree, Logistic Regression, Naive Bayes, and Support Vector Machines to these numerical representations. This step aimed to uncover valuable insights and patterns hidden within the text data.

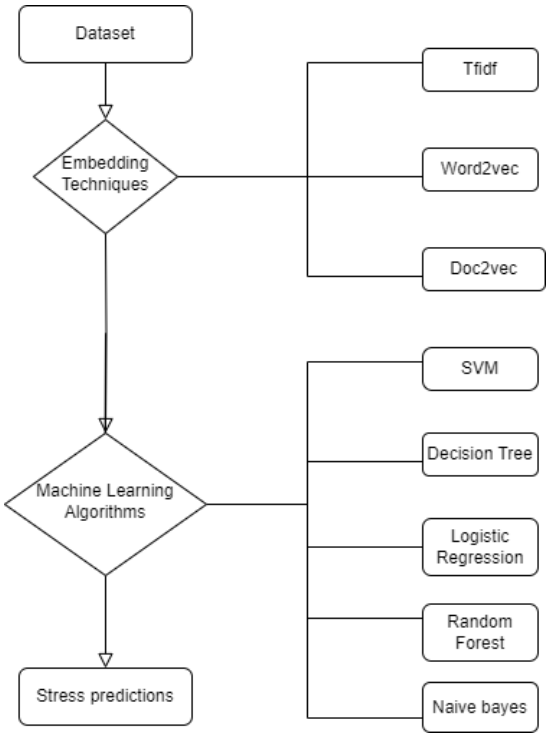

Fig 1: Proposed System Workflow

A. Dataset Used

The dataset provided contains text samples in two languages: Tamil and Telugu. In the Tamil dataset, there are 1,784 samples labeled as "Stressed" and 3,720 samples labeled as "Non-Stressed" for training, out of a total of 5,504 samples. The test dataset for Tamil comprises 1,020 samples. Similarly, in the Telugu dataset, there are 1,783 "Stressed" samples and 3,314 "Non-Stressed" samples for training, out of a total of 5,097 samples. The test dataset for Telugu includes 1,050 samples. Additionally, a development dataset is provided to assess the model's performance.

Table 1: Tamil dataset Description

Dataset	No. of Comments
Train	5504
Test	1020

Table 2: Telugu dataset Description

Dataset	No. of Comments
Train	5097
Test	1050

B. Embedding Techniques

TF-IDF: TF-IDF plays a crucial role in NLP by assessing the significance of terms within documents based on their frequency in the document and rarity across the corpus. This method is vital for tasks like document classification, sentiment analysis, and information retrieval, enabling accurate analysis of textual data across diverse industries and applications.

Word2vec: Word2Vec is a widely utilized technique for generating word embeddings in NLP. It represents words as dense vectors in a high-dimensional space, capturing semantic relationships between words. Through training on extensive text datasets, Word2Vec effectively captures contextual information, facilitating language modeling, document classification, and word similarity computation.

Doc2vec: Doc2Vec extends the principles of Word2Vec to entire documents, generating document embeddings. By representing documents as dense vectors in a high-dimensional space and training on large document collections, Doc2Vec captures the unique features and context of each document. This enables tasks such as document similarity assessment, sentiment analysis, and document classification, contributing significantly to various NLP applications.

C. Machine Learning algorithms

Logistic Regression: For binary classification issues, logistic regression is a predictive modeling technique that is applied. Using a logistic function that limits the output to a range between 0 and 1, it predicts the likelihood that an instance belongs to a specific class. In order to optimize the likelihood of the observed data, the algorithm is trained to identify the ideal parameters.

Decision Trees: Decision trees are non-parametric supervised learning methods suitable for regression and classification applications. An internal node, a leaf node, branches and a root node make up the hierarchical tree structure that the method builds. A final prediction is made at the leaf nodes after decisions are made at each node based on feature values. Decision trees can represent intricate relationships in the data and are interpretable.

Naïve Bayes: A popular probabilistic classifier for text classification is Naïve Bayes. It is a member of the generative learning algorithm family and models the input distribution for

a particular class. To make computations simpler, the approach assumes that features are conditionally independent. Naïve Bayes is an efficient method for handling huge datasets and high-dimensional feature spaces because it uses the Bayes Theorem to determine the likelihood of a class given observed features.

Random Forest: An ensemble learning system called Random Forest integrates the results of several decision trees. By adding feature randomness, it improves on the bagging method and guarantees low correlation between the trees. Regression and classification issues are easily handled by Random Forest, which also provides robustness against overfitting and enhances generalization performance. It increases prediction accuracy and variety by taking into account subsets of features.

Support Vector Machine: Strong supervised learning models for regression, outlier detection, and classification are Support Vector Machines (SVM). SVMs function by determining the best hyperplane in the feature space to divide various classes. The approach maximizes the margin between classes by transforming the data into a higher-dimensional space. Because SVMs can handle complex correlations in the data, they are useful in many disciplines, such as image identification, signal processing, natural language processing, and healthcare.

IV. Results And Discussion

Table 3 shows the performance of various machine learning algorithms for Telugu dataset It seems like Random Forest consistently performs well across all three embedding methods, with accuracies of 0.99 for TF-IDF, Doc2Vec, and Word2Vec. SVM and Logistic Regression also show high accuracies, particularly with TF-IDF and Word2Vec embeddings. Decision Tree achieves relatively high accuracies as well, with 0.98 across all embedding methods. However, Naive Bayes shows lower accuracies compared to other algorithms, particularly with Doc2Vec and Word2Vec embeddings.

Table 3. Accuracy scores for Telugu dataset

Classifiers	TF-IDF	Doc2Vec	Word2Vec
SVM	0.99	0.98	0.97
Decision Tree	0.98	0.98	0.98
Naïve Bayes	0.93	0.65	0.65
Logistic Regression	0.99	0.97	0.97
Random Forest	0/99	0.99	0.99

Table 4 shows the performance of various machine learning algorithms for Tamil dataset. Random Forest shows strong performance, particularly with Word2Vec embedding, achieving accuracies of 0.98, 0.97 and 0.99 across TF-IDF, Doc2Vec, and Word2Vec, respectively. Support Vector Machine (SVM) achieves consistent accuracies of 0.98 across TF-IDF, Doc2Vec and Word2Vec embeddings.

Table 4. Accuracy scores for Tamil dataset

Classifiers	TF-IDF	Doc2Vec	Word2Vec
SVM	0.98	0.98	0.98
Decision Tree	0.97	0.92	0.97
Naïve Bayes	0.92	0.92	0.91
Logistic Regression	0.96	0.95	0.97
Random Forest	0/98	0.97	0.99

Figures 2 to 11 present the confusion matrices obtained with Word2Vec embedding for Telugu and Tamil datasets. It is observed that Random Forest outperforms other classifiers with all the three embedding techniques for both the datasets. Among the embedding techniques used, it appears that Word2Vec embedding generally outperforms TF-IDF and Doc2Vec embeddings, particularly when used with Random Forest and Logistic Regression classifiers which achieved the highest accuracies with Word2Vec.

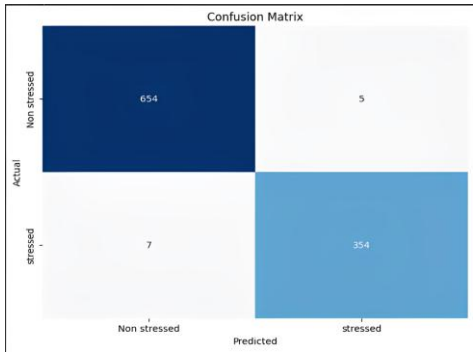

Fig. 2. Confusion Matrix of SVM for Telugu dataset

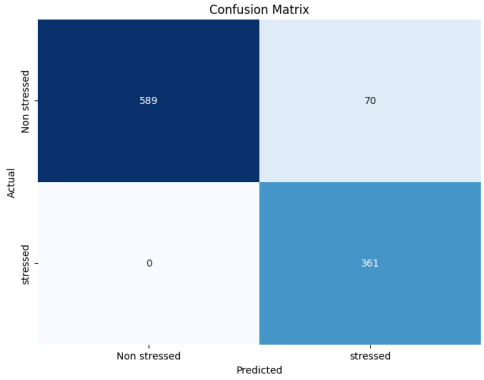

Fig. 3. Confusion Matrix of Naïve Bayes for Telugu dataset

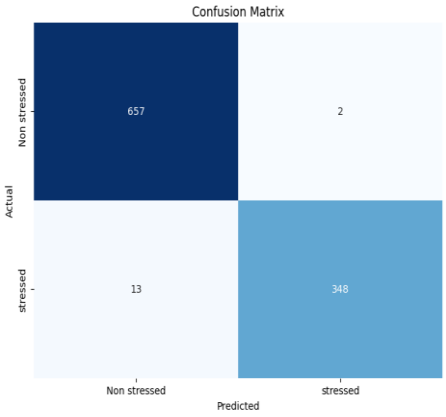

Fig. 4. Confusion Matrix of Logistic Regression for Telugu dataset

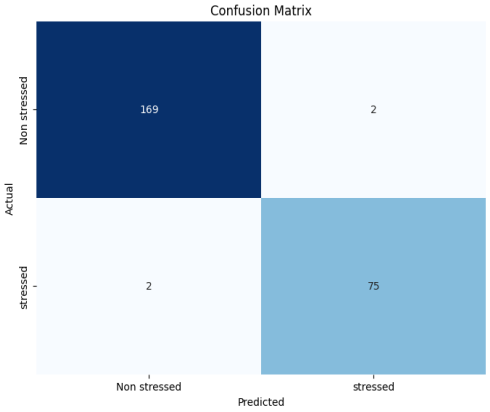

Fig. 5. Confusion Matrix of Decision Tree for Telugu dataset

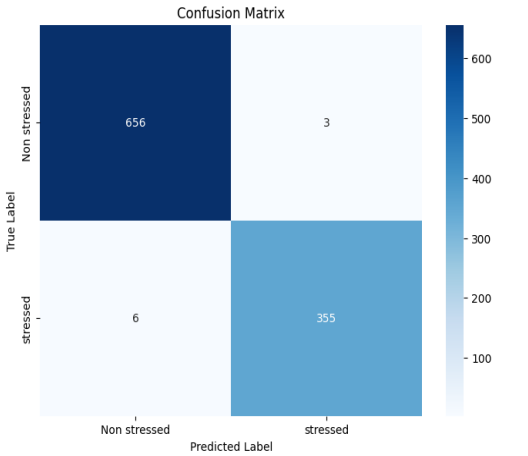

Fig. 6. Confusion Matrix of Random Forest for Telugu dataset

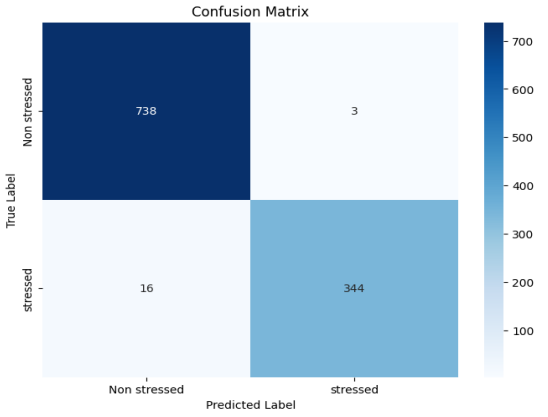

Fig. 7. Confusion Matrix of SVM for Tamil dataset

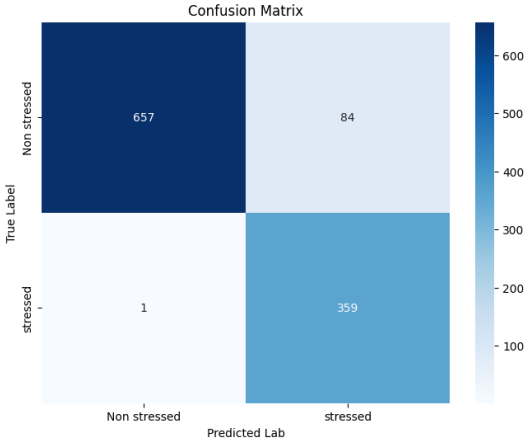

Fig. 8. Confusion Matrix of Naïve Bayes for Tamil dataset

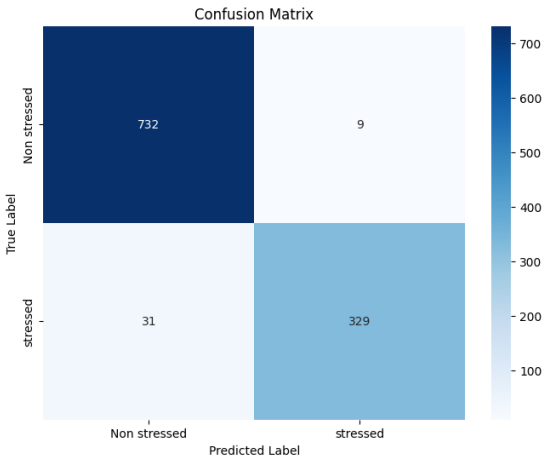

Fig. 9. Confusion Matrix of Logistic regression for Tamil dataset

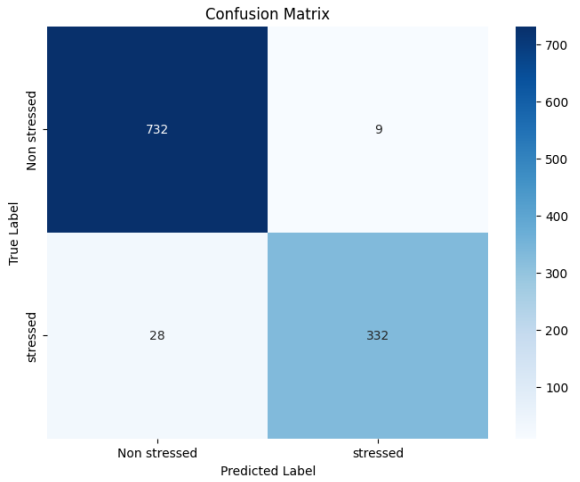

Fig. 7. Confusion Matrix of Decision Tree for Tamil dataset

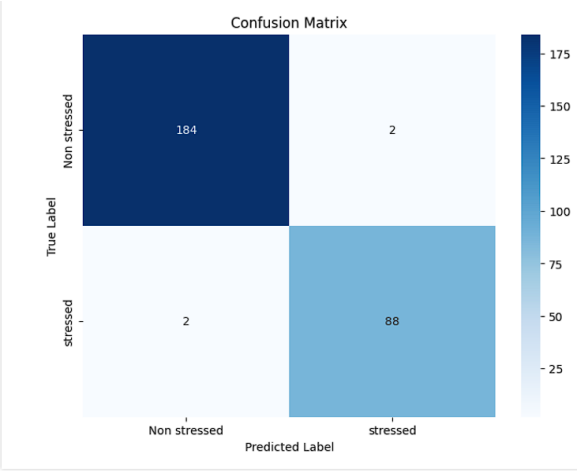

Fig. 11. Confusion Matrix of Random Forest for Tamil dataset

V. Conclusion

This study underscores the efficacy of employing various embedding techniques and machine learning algorithms for stress analysis in Tamil and Telugu data. Leveraging datasets sourced from DravidanLangTech-2024, we demonstrated the effectiveness of TF-IDF, Doc2Vec and Word2Vec embeddings in representing textual information numerically. Through the utilization of Random Forest, SVM, Decision Tree, Naïve Bayes and Logistic Regression algorithms, we successfully detected indicators of stress with notable accuracies across both Tamil and Telugu datasets. From the simulations done, Random Forest model with Word2Vec embedding produced a remarkable accuracy of 99% for both Tamil and Telugu datasets.

References

[1] Gowtham, B., et al. "Stress Analysis Using Machine Learning." *Applied Computing for Software and Smart Systems: Proceedings of ACSS 2022*. Singapore: Springer Nature Singapore, 2023. 227-234.

[2] Alwin Infant, P., et al. "Stress Analysis Prediction for Coma Patient Using Machine Learning." *International Conference on Data & Information Sciences*. Singapore: Springer Nature Singapore, 2023.

[3] Ratul, Ishrak Jahan, et al. "Analyzing perceived psychological and social stress of university students: A machine learning approach." *Heliyon* 9.6 (2023).

[4] Rahman, Ahnaf Akif, et al. "Perceived stress analysis of undergraduate students during covid-19: a machine learning approach." *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2022.

[5] Badr, Yara, et al. "Mental Stress Detection and Mitigation using Machine Learning and Binaural Beat Stimulation." *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023.

[6] Oduntan, Adenrele, et al. ""'I Let Depression and Anxiety Drown Me…": Identifying Factors Associated With Resilience Based on Journaling Using Machine Learning and Thematic Analysis." *IEEE journal of biomedical and health informatics* 26.7 (2022): 3397-3408.

[7] Manjunath, Prakruthi, et al. "Predictive analysis of student stress level using Machine Learning." *Int. J. Eng. Res. Technol* 9 (2021): 76-80.

[8] Swain, Satyananda, and Manas Ranjan Patra. "Analysis of Depression, Anxiety, and Stress Chaos Among Children and Adolescents Using Machine Learning Algorithms." *International Conference on Innovations in Intelligent Computing and Communications*. Cham: Springer International Publishing, 2022.

[9] Najjar, Esraa, and Aqeel Majeed Breesam. "Supervised Machine Learning a Brief Survey of Approaches." *Al-Iraqia Journal for Scientific Engineering Research* 2.4 (2023): 71-82.

[10] Rahman, Ahnaf Akif, et al. "Perceived stress analysis of undergraduate students during covid-19: a machine learning approach." *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*. IEEE, 2022.

[11] Zainudin, Z., et al. "Stress detection using machine learning and deep learning." *Journal of Physics: Conference Series*. Vol. 1997. No. 1. IOP Publishing, 2021.

[12] Walambe, Rahee, et al. "Employing multimodal machine learning for stress detection." *Journal of Healthcare Engineering* 2021 (2021): 1-12.

[13] Rosales, Marife A., et al. "Physiological-based smart stress detector using machine learning algorithms." *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. IEEE, 2019.

[14] Archana, V. R., and B. M. Devaraju. "Stress detection using machine learning algorithms." *International Journal of Research in Engineering, Science and Management* 3.8 (2020): 251-256.

[15] Reddy, U. Srinivasulu, Aditya Vivek Thota, and A. Dharun. "Machine learning techniques for stress prediction in working employees." *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2018.