Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Multimodal Framework for Stress Detection Using Textual and Behavioural Signals With Adaptive Fusion

Adhith A P
Dept. of Computing Technologies SRM University
Kattankulathur, India

Shorya Vardhan Tyagi
Dept. of Computing Technologies SRM University
Kattankulathur, India

Mr. Ganesh Shanker S
Dept. of Computing Technologies SRM University
Kattankulathur, India

***Abstract*—Stress detection proves to be an important problem in modern digital frameworks where behavioural and textual signals provide useful inferences. The proposed work suggests a multimodal framework that combines the textual and behavioural signals to detect a continuous index that allows to infer the stress levels. The system combines a BERT architecture that is fine-tuned for textual signals along with a neural network framework for behavioural signals like sleep hours, sleep quality, physical activity levels, heart rate and step count. For integration, an adaptive fusion mechanism is employed, which dynamically assigns the weights of the input sources. During the experimentation phase, the dataset characteristics greatly influenced the model behaviour. In some cases, highly stress-specific datasets resulted in heavily biased models, while the general emotion-based datasets proved to be more stable but less expressive. The final system achieved stable and interpretable results, represented using various performance metrics such as AUC, Precision, Recall and F1-score. The results highlight the importance of dataset alignment in multimodal frameworks, and the behavioural signals take dominance when the textual signals lack sufficient variance and expressiveness.**

Keywords—Stress Detection, Multi-Modal Learning, Transformer Models, Behavioural Analysis, Mental Health, Fusion Model

I. INTRODUCTION

Stress is a critical factor that affects the human body mentally and physically. With the modern-day lifestyle and work conditions, there arises a need for a system that can automatically measure the stress levels using available data. Stress builds up over time due to continuously being subjected to harsh physical and mental situations; thus, early identification of lifestyle changes can prove to be crucial.

Stress is not just a physiological response, but also it is formed based on how different individuals perceive their situation or environment and their response to it. According to the transactional model of stress proposed by Richard Lazarus and Susan Folkman[1], stress takes shape from the interaction of an individual with their environment. This model emphasises that stress is not just a result of external physiological factors but is also formed by an individual's internal perception and emotional response. Therefore, both behavioural and subjective expressions play an important role in evaluating and understanding stress.

Traditional methods to detect stress typically include a singular modality. Since each modality has its own limitations, relying on a singular modality could be inefficient. Textual data can capture personal, emotional and contextual information but if the content does not prove to be related to stress it becomes unreliable, whereas, in the case of behavioural data consists of objective and precise data but may lack contextual depth behind it.

Due to their complementary strengths and limitations, there has been a growing interest in implementing multimodal approaches by combining multiple input sources. Such systems aim to provide a more comprehensive view of stress analysis, taking into account both objective and subjective expressions.

In this work, we propose a multimodal stress detection system using textual and behavioural signals with an adaptive fusion system. For textual inputs, we implement a fine - tuned BERT architecture that allows the comprehension of subjective expressions, while in the case of behavioural inputs like sleep hours, sleep quality, physical activity levels, heart rate and step count using a neural network system. The two modalities are integrated using a fusion strategy that dynamically allocates the weights based on relative confidence.

A key focus point of this work is not only the multimodal system but also the exploration of how different dataset characteristics affect the respective model behaviours. During experimentation, multiple datasets were analysed, including a stress-specific dataset as well as a general emotions-based dataset. While analysing a stress-specific dataset, it was observed that even though the model was providing strong signals, the model turned out to be heavily biased. In contrast, the general emotion dataset provided a more stable result but with little variance in the model prediction. Thus proving the importance of dataset selection and alignment in multimodal systems.

Another key inference of this study is the examination of modality contribution. By determining how much each

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

output of these modalities contributes to the final prediction, we gain insight into the relative importance of behavioural and textual features. The work shows that the behavioural features tend to dominate when the textual data indicates lack of sufficient variability, emphasising the role of structural data in this context.

Overall, the work contributes to the understanding of how different system designs, dataset characteristics and fusion strategies affect the performance. Instead of prioritising performance alone, the study emphasises interpretability, practicality and stability. This study enables us to deepen the understanding of multimodal systems and how to make them more balanced and robust.

II. LITERATURE REVIEW

Stress detection has been widely analysed using a plethora of data sources and computational approaches. Traditional methods consisted of processing physiological and behavioural data like heart rate, sleep pattern or physical activity levels[2]. These metrics are key to understanding the stress levels directly in the human body and are objective data sources. While they provide objective data, they lack the contextual and emotional nuances that paint the whole picture.

With the advancement of natural language processing, stress detection using text-based input has gained significant attention. Social media platforms and user-generated content have been widely used for emotional and physiological state analysis[4]. Transformer-based models, like BERT, have gained popularity for their strong understanding of languages and understanding subtle emotional patterns[3]. However, the effectiveness of these models depends entirely on the quality of the datasets, and many such datasets provide ambiguous or neutral content that limits the ability of these models to understand the emotional patterns.

In order to combat the limitation put forward by singular modality systems, recent systems have explored multimodal systems that take into account multiple sources of data. These systems integrate both behavioural and textual data signals to improve prediction accuracy and robustness[5]. These multimodal framework works implement fusion strategies that typically range from simple weighted averages to complex attention-based mechanisms[5]. These systems have been shown to capture both subjective and objective aspects of stress.

Despite the advancements, several limitations remain. One key challenge is dataset alignment for these multimodal systems, where multiple modalities may not align with similar stress indicators. In this work, we build upon the existing multimodal systems by not only implementing the framework but also analysing how dataset characteristics affect modality contribution and the overall model behaviour. This study provides insight into the practical challenges of deploying multimodal systems.

III. METHODOLOGY

A. System Overview

The proposed system comprises a multimodal framework for stress detection that leverages text and behavioural signals. The system has a parallel framework of a BERT transformer-based model that processes textual inputs and a neural network to process behavioural signals. This model, instead of relying on a singular type of data, processes two types in parallel to capture both subjective and objective aspects of stress.

At a higher level, the system contains two pipelines that compute their respective scores. Then these scores are combined through an adaptive fusion strategy, which determines the modality contribution of the two models dynamically. The fusion outputs a continuous score from 0 to 1, which is then categorised into discrete levels for easier interpretability.

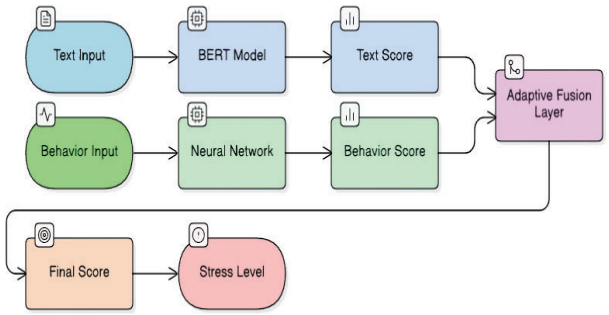

Figure 1: System Architecture

B. Text-Based Module

The textual pipeline of the system is processed using a pre-trained BERT model, which is fine-tuned for the task of stress estimation via text signals. BERT was chosen for its ability to capture contextual relationships within textual data, thus making it effective for analysing subjective and psychological content.

The text input is passed using a tokeniser, which converts the raw textual data into tokens that can be understood by the BERT model. These tokens are then passed through the transformer encoder, from where contextual embeddings are generated based on the relationships between the words in the text. The stress score produced by the encoder is then passed through a classification layer, which is then followed by a sigmoid function that represents the score as a value between 0 and 1. This score represents the likelihood of stress inferred from the textual input.

During experimentation, it was observed that the textual model tends to produce relatively stable outputs with limited variation. This behaviour can be attributed to the nature of the dataset, where not all samples strongly reflect stress-related content.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

C. Behaviour-Based Module

The behaviour-based model processes structured data of features that directly convey the objective physical state of the individual. The features include sleep hours, subjective sleep quality, physical activity level (number of minutes/day), Heart rate level and daily step count. Before being fed into the model, the features are normalised to ensure stability and improve the performance of the neural network. These normalised inputs are then passed on to a feedforward neural network consisting of two fully connected hidden layers with 32 and 16 neurons, respectively. Each layer learns patterns and relationships between the behavioural indicators and stress levels, enabling the model to capture underlying trends in the data. The output score by the behaviour model is a continuous score of 0 to 1. Compared to the textual model, this component generally produces more varied outputs, indicating a stronger correlation between behavioural data and stress.

D. Fusion Layer

To integrate the outputs of the textual and behavioural models, an adaptive fusion mechanism is employed. Rather than assigning fixed weights to each modality, the system dynamically determines their contribution based on their respective outputs.

Let the stress scores from the text and behavioural models be denoted as S_t and S_b, respectively. The fusion weight is computed as:

$$\alpha = S\_t / (S\_t + S\_b)$$

Using this weight, the final stress score is calculated as:

$$S\_f = \alpha S\_t + (1 - \alpha) S\_b$$

This approach allows the system to adjust the influence of each modality depending on its relative confidence. In practice, it was observed that the behavioural component often contributes more significantly due to higher variability in its predictions, while the textual component provides a stabilising effect.

E. Stress Classification

The final output of the system is a continuous stress score, which is mapped into discrete categories to enhance interpretability. The continuous stress score is categorised into three levels using empirically selected thresholds: low stress (0–0.4), moderate stress (0.4–0.7), and high stress (0.7–1.0). The thresholds are selected based on the distribution of predicted scores to ensure meaningful separation between categories.

IV. Results And Analysis

A. Evaluation Metrics

To evaluate the performance of the proposed system, a set of standard classification metrics is used. These include accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC). Each of these metrics provides a different perspective on model performance. While accuracy gives an overall measure of correctness, precision and recall help in understanding how well the model identifies stress-related instances. The F1-score provides a balance between precision and recall, and AUC reflects the model's ability to distinguish between different stress levels. Together, these metrics provide a detailed and balanced evaluation of the system, enabling a clear comparison between the textual, behavioral, and fusion models.

B. Quantitative Analysis

TABLE I. **Performance Comparison of Models**

Model	Accuracy	Precision	Recall	F1 Score	AUC
BERT (Text)	0.43	0.39	0.43	0.35	0.55
Behavioural	0.41	0.50	0.41	0.38	0.99
Fusion	0.71	0.77	0.71	0.72	0.94

From the results, it is evident that the fusion model outperforms both individual models across all major evaluation metrics.

C. Graphical Analysis

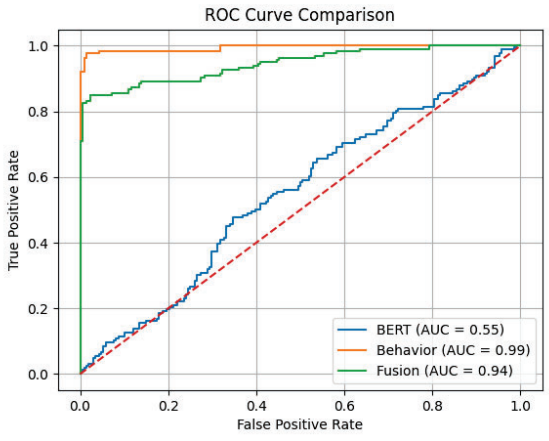

Figure 2:ROC Curve of all three models

The Receiver Operating Characteristic (ROC) curve comparison for the textual, behavioural, and fusion models is shown in figure. The behavioural model achieves an AUC value close to 0.99, indicating a strong ability to distinguish

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

between stress levels. The fusion model also demonstrates high performance with an AUC of approximately 0.94, outperforming the textual model significantly. In contrast, the textual model exhibits a relatively lower AUC of around 0.55, suggesting limited discriminative capability. This comparison highlights the effectiveness of combining modalities, as the fusion model maintains strong separability while improving overall classification performance. textual, behavioural, and fusion models.

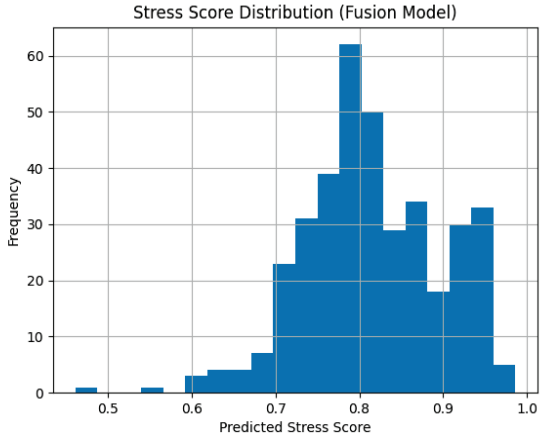

Figure 2:Stress Score Distribution for the fusion model

The distribution of predicted stress scores for the fusion model is illustrated in the figure. The histogram shows that the predicted values are spread across a wide range, rather than being concentrated at a single point. This indicates that the model produces sufficiently varied outputs, which is essential for meaningful classification. The spread of values also justifies the use of percentile-based thresholding, as it enables a balanced division of predictions into low, moderate, and high stress categories.

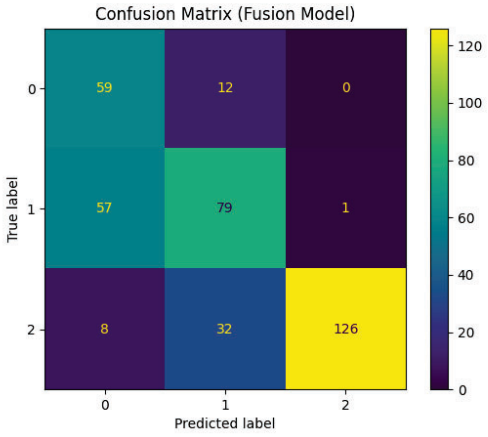

Figure 3:Confusion Matrix for the fusion model

The confusion matrix for the fusion model is presented in figure. It can be observed that the model performs well in identifying high-stress instances, with a large number of correct predictions in the highest stress category. However, some misclassification occurs between adjacent classes, particularly between low and moderate stress levels. This suggests that while the model is effective in distinguishing extreme cases, borderline instances remain challenging. Overall, the confusion matrix supports the improved classification capability of the fusion model.

D. Model Wise Analysis

The textual model, based on BERT, achieves moderate performance with an accuracy of approximately 43% and an AUC close to 0.55. This indicates that while the model is able to capture some contextual and emotional cues from text, it lacks strong discriminative power for stress detection. This limitation can be attributed to the nature of the dataset, which is not explicitly designed for stress classification.

The behavioural model demonstrates a different performance pattern. Although its accuracy and F1-score are relatively moderate, the model achieves a very high AUC value of 0.99. This implies that the model is highly effective at ranking and separating stress levels based on physiological and lifestyle features such as sleep duration, heart rate, and physical activity. However, the conversion of continuous outputs into discrete categories using thresholds affects their classification performance.

The fusion model achieves the best overall performance, with an accuracy of approximately 71% and an F1-score of 0.72. This improvement highlights the effectiveness of combining textual and behavioural modalities. The textual model contributes contextual and emotional information, while the behavioural model provides strong physiological indicators of stress.

By integrating these complementary sources of information, the fusion model is able to overcome the individual limitations of each modality. The resulting predictions are more balanced and robust, leading to improved classification performance.

E. Dataset Impact and Observations

The experimental results clearly indicate that dataset characteristics have a significant influence on model performance. During the initial phase, a Reddit-based mental health dataset was explored for the textual modality. This dataset predominantly contained posts reflecting high levels of stress, resulting in a strong bias and limited variability in the data. Consequently, the textual model showed poor discriminative capability, with performance close to random levels (AUC $\approx$ 0.55), indicating difficulty in distinguishing between different stress levels.

To address this issue, a more balanced emotion-based dataset was used for training the textual model. This dataset provided a wider range of emotional expressions, which improved the stability of predictions. However, since emotional labels do not directly correspond to stress levels, the model's ability to capture stress-specific patterns remained limited. This is reflected in the moderate performance of the textual model,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

with an accuracy of approximately 43% and an F1-score of 0.35.

In contrast, the behavioral dataset consists of structured features such as sleep duration, physical activity, heart rate, and daily step count, which are more directly related to stress. This strong correlation is evident in the behavioral model's high discriminative performance, achieving an AUC value close to 0.99. However, despite this high separability, the classification accuracy remains moderate (approximately 41%), primarily due to the use of threshold-based categorization of continuous outputs.

The fusion model combines the strengths of both modalities, integrating contextual information from textual data with physiological indicators from behavioural features. This results in a significant improvement in overall performance, with the fusion model achieving an accuracy of approximately 71%, an F1-score of 0.72, and an AUC of 0.94. These results demonstrate that while individual modalities capture partial aspects of stress, their combination leads to a more robust and effective system.

F. Key Observations

- The behavioural model provides strong discriminative capability due to the direct relationship between features and stress levels.
- The textual model alone is insufficient for accurate stress prediction due to the lack of stress-specific labelling.
- The fusion approach effectively combines complementary information from both modalities, resulting in improved performance.
- The choice and characteristics of datasets play a crucial role in determining model behaviour and overall system effectiveness.

V. Conclusion And Future Work

A. Conclusion

This study presents a multimodal framework for stress detection by integrating textual and behavioural signals. The results demonstrate that individual modalities capture different aspects of stress, but are limited when used independently. Thus, the introduction of a multimodal system consisting of a textual and behavioural model. The textual model, trained on a general emotion-based dataset, achieved moderate performance with an accuracy of approximately 43% and an AUC of 0.55, indicating limited capability in identifying stress-specific patterns. Similarly, the behavioural model showed strong discriminative ability with an AUC of 0.99, but its classification performance remained moderate due to threshold-based categorisation.

The fusion model, which combines both modalities through an adaptive weighting mechanism, achieved the best overall performance with an accuracy of approximately 71%, an F1-score of 0.72, and an AUC of 0.94. This demonstrates that integrating contextual information from text with physiological indicators significantly improves prediction accuracy and robustness. The findings highlight that multimodal approaches are more effective than single-modality models for stress detection, particularly when the modalities provide complementary information.

The study also emphasises the importance of dataset characteristics in determining model performance. It was observed that highly biased datasets reduce variability and affect prediction quality, while more balanced datasets improve stability but may lack task-specific signals. Therefore, careful dataset selection and alignment are crucial for achieving reliable results in multimodal systems.

B. Future Work

Several directions can be explored to further improve the proposed system. One key area is the use of more appropriate and stress-specific textual datasets. The current results indicate that the performance of the textual model is limited by the lack of direct alignment between emotional expressions and stress levels. Incorporating datasets explicitly annotated for stress detection can significantly improve the effectiveness of the BERT-based model.

Another important direction is the development of more advanced fusion strategies. While the current adaptive weighting approach provides a simple and interpretable mechanism, more sophisticated techniques such as attention-based fusion or learnable weighting schemes could better capture the complex relationships between modalities and further improve performance.

From a practical perspective, the system can be extended by integrating real-world data sources. This includes incorporating continuous behavioural signals from wearable devices such as smartwatches and fitness trackers, enabling real-time stress monitoring. Such an implementation would improve the applicability of the system in healthcare and personal wellness domains.

Additionally, future work can focus on constructing or utilising a unified multimodal dataset where both textual and behavioural data correspond to the same individuals. This would ensure better alignment between modalities and allow for more accurate and realistic evaluation of the system.

Finally, improvements in model architecture, including fine-tuning of hyperparameters and the use of more advanced transformer-based models, can further enhance prediction accuracy. These advancements would contribute to building a more robust, scalable, and practical stress detection system.

References

[1] R. S. Lazarus and S. Folkman, Stress, Appraisal, and Coping. New York, NY, USA: Springer, 1984.

[2] T. Gjoreski, M. Gams, H. Gjoreski, M. Luštrek, and M. Gradišek, "Continuous stress detection using a wrist device: In laboratory and real life," in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., 2016, pp. 1185–1193.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[4] S. Calvo, D. Milne, M. Hussain, and H. Christensen, “Natural language processing in mental health applications using non-clinical texts,” Natural Language Engineering, vol. 23, no. 5, pp. 649–685, 2017.

[5] S. D'Mello and A. Graesser, “Multimodal affect detection,” in Proc. Int. Conf. Multimodal Interfaces, 2010, pp. 325–328.

[6] D. D mszky, D. Movshovitz-Attias, J. Ko, G. Cowen, G. Nemade, and S.Ravi, “GoEmotions: A Dataset of Fine-Grained Emotions,” in Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4040–4054.

[7] “ eddit Mental Health Dataset,” Kaggle. [Online]. Available: https://www.kaggle.com/

[8] “Sleep Health and Lifestyle Dataset,” Kaggle. [Online]. Available: https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset