AI Based Stress and Mental Health Monitoring System Using Chatbot, Speech and Facial Analysis

Er. Ria Mathews, Nandana Anil, Jisa Elsa Sam, Naivedya S Krishna, Smera Sara Kurian
Department of Computer Science and Engineering
Saintgits College of Engineering
Kottayam, Kerala, India

***Abstract*—The increase in academic pressure and workplace stress has led to higher levels of mental illness, including stress, anxiety, and depression. When stress is not handled and recognized early, there can be significant health issues, both mentally and physically. Traditional ways of evaluating mental health rely on self-reporting and clinical interviews, which can be very subjective and time-consuming and not always able to be monitored continuously or in real-time. Because of this, artificial intelligence (AI) has emerged as a viable option for providing automated assessments of mental health through technology. This paper discusses the current state of AI-based monitoring of mental health and stress and provides an overview of studies that have examined AI-based chatbots, audio signal analysis, and facial recognition. This systematic review of the literature also provides a summary of various machine learning and deep learning techniques used to detect patterns of stress using multimodal data. The conclusions drawn from this study suggest that using multiple data sources together improves the accuracy and robustness of AI-based systems compared to single-modality systems, making them more suitable for practical use in mental health settings.**

***Index Terms*—Mental health monitoring, Stress detection, Anxiety and depression, Artificial intelligence (AI), Machine learning, Deep learning, Multimodal data, AI-based chatbots, Audio signal analysis, Facial recognition, Automated mental health assessment, Stress classification, Pattern recognition, Healthcare analytics**

I. Introduction

Emotional stability, cognitive function, and overall quality-of-life are all greatly impacted by an individual's mental health. Stress is a leading cause of the mental health issues that affect people of all ages, and if not addressed in a timely manner, may lead to the development of conditions such as anxiety, depression, sleep disorders, or heart disease. Although stress can have serious implications for an individual's mental health, many people are unaware that they are suffering from it because of a lack of understanding, social stigma associated with mental health, and lack of access to professional support services.

Some of the common methods that are used to assess stress include psychological questionnaires, face-to-face interviews, and a clinical assessment. These types of assessments can be helpful, but they rely heavily on a lot of subjective opinions and can be subject to a lot of variation and bias. Furthermore, these approaches do not provide a means for continuous or real-time assessment of stress. With the improvement of technology such as Artificial Intelligence (AI) and machine learning, new avenues for non-invasive, automated mental health assessments have become available. AI-enabled systems now have the capability to evaluate emotional, behavioural, and physical changes to provide data to identify potential signs of stress and provide treatment.

Chatbots are a source for conversational research, speech analysis uses vocal characteristics to find stress and facial analysis helps detect emotional deviations. Gathering data in this way provides a fuller picture of an individual's mental status.

Typically, one source of data (like chatbots) encounters problems due to environment-based noise, signals or variations among people. By combining chatbot conversations, vocal characteristics and facial characteristics to create a multi-modal AI-based stress monitoring system, we can solve these challenges. Using multiple sources of data to measure stress enhances the reliability of results, reduces the likelihood of wrong predictions and increases overall accuracy.

Thus, there is a great deal of potential for multi-modal AI-based stress-identifying systems to identify stress before it occurs, monitor stress continuously and provide personalised support. They act as a means of support for mental health professionals, providing evidence based data for real time assessment, thereby improving overall quality of mental health care and optimising preventative measures.

II. Related Work

This study has provided an overview of recent innovations in the use of artificial intelligence-based systems for monitoring psychological well-being and stress. With the advent of chatbots and other forms of AI technology, users now have access to continuous, automatic, and passive assessments of their mental state. By combining three kinds of data—text, voice, and visual—this multi-modal analysis allows for a better understanding of how someone emotionally reacts to stress than could be achieved by relying on any single type of data alone.

Using multiple methods for assessment of emotional state yields significantly better outcomes than using just a single method. A chatbot can be used to collect contextual and behavioural data; vocal analysis can identify very small changes

indicative of increased stress; and video analysis can assess how someone is feeling emotionally, without necessarily needing to voice their feelings verbally. The combination of these methods provides a greater sense of certainty and increases the chance of successful prediction of emotional state.

The ability of AI systems to monitor individuals for the early identification of stress and to provide ongoing real-time objective data may improve the mental health intervention decision-making process and reduce the burden placed on mental health professionals. Accessibility and confidentiality provided to users may also assist this type of AI system to promote early seeking of mental health assistance by the general public without the fear of being judged by society.

The deployment of these systems in practice must be done carefully considering the following challenges: Data security, Ethical and Potential Bias in Learning Models. Meeting these challenges is crucial to ensuring the responsible and reliable use of AI. As Artificial Intelligence continues to develop and explainability and security practices advance, the use of AI-assisted mental health monitoring will become a central part of Future Healthcare Models, providing better Preventive Care and Improved Mental Health Outcomes.

III. METHODOLOGY

The aim of this research was to develop a multimodal AI framework for monitoring mental health and stress levels using a chatbot interface to integrate facial expression analysis, speech signal processing, and conversation text analysis. Machine learning, deep learning, and Natural Language Processing (NLP) will be employed in order to create an unobtrusive and automated mental health assessment system that will provide users with continuous assessment of their mental health.

A modular/multi-step architectural approach was utilized in order to improve adaptability, maintain the user's privacy while utilizing their data to accurately detect stress, and provide a more accurate system with higher reliability.

A. Data Collection

The AI Framework will be able to process several different types of user input via three independent and parallel "pipelines."

- **Facial Expression Data-** Facial samples will be gathered from the users' own recorded videos, as well as photographs. The collected dataset will include samples of a range of emotions in various light settings, as well as samples presented at varied angles to better simulate the user's environment.
- **Speech Data-** The user will provide an audio recording of their voice for analysis to assess voice activation responses to stress. Each of these recordings will vary in terms of pitch, tone, rate of speech, and emotional intensity; these types of variations in voice activation are indicators of stress.
- **Conversational Text Data-** The AI Framework will use the user's verbal statements made through the chatbot to develop a conversational dataset, which will serve as the basis for stress analysis to be provided by the AI Framework. The data contained in the conversational dataset will include both stressed-related phrases and patterns of sentiment analysis, as well as language of emotion.

All data collected through this system is used only temporarily, processed according to the ethical standards set forth in this document, to allow for complete user confidentiality and protection of the data.

B. Data Preprocessing

The Modality Workflow for Data Preprocessing is specific to each type of input (Modality). In the case of Video inputs, the following steps are performed:

1) ***Facial Modality Data Preprocessing****:* The video input is cut into a series of frames at specific times.

The facial areas are extracted from the frames using face detection and alignment methods.

Next, the extracted facial images are resized and Normalized before using them in the Models.

2) ***Speech Modality Data Preprocessing****:* For Audio input, the audio sample is converted to a fixed sampling rate.

Noise suppression techniques are applied to remove or minify unwanted background noise from the Audio as well as improve audio quality for the intended user.

Mel-Frequency Cepstral Coefficients (MFCCs) represent the vocal characteristics of Voice (Pitch and Energy).

3) ***Text Modality Data Preprocessing****:* For Text input, the data is cleaned up by Tokenizing, removing stop-words and transforming to Leurlemmas.

Using Natural Language Processing (NLP) methods, Emotion keywords and Sentiment Polarity will be a major area of focus.

The Text input will ultimately be transformed into Numerical Vector Columns (Numerical Embedding) for use with Machine Learning models.

C. Design of System

The structure of the system consists of multiple modular architectures built on top of Deep Learning algorithms. For each type of data analyzed, an independent module will analyze the data first; after each independent analysis is performed, the data will then be aggregated prior to coming to the final decision. The modular structure allows for flexibility of the system. Each module allows for the addition or extension of new modules to capture additional aspects of stress detection or increase the accuracy of the individual module.

1) ***Recognition of Emotions through Facial Recognition****:* For emotion detection via facial recognition, a deep learning algorithm based on ResMaskNet is used. The ResMaskNet deep learning algorithm uses both attention and residual connections. This enables the ResMaskNet to focus on identifying morphological traits that represent emotional states. The ResMaskNet deep learning algorithm can identify emotions including (but not limited to): happiness, sadness, rage, fear,

surprise, disgust, or neutral emotion, as well as generate a probability score for each identified emotion.

2) ***Detecting Stress from Speech****:* Detection of speech-based stress is accomplished through the use of a deep neural network designed in TensorFlow and Keras. For speech-based detection, the model extracts Mel-frequency cepstrum coefficients (MFCCs) from audio files to determine differences in voice pitch, tone, and energy. Based on the characteristics associated with the characteristics of the voice, the model classifies the stressful speech condition vs. non-stressed speech condition.

Our system analyzes user conversations via a chatbot using Natural Language Processing and Sentiment Analysis. Through analyzing the emotional tone, stress vocabulary, and syntax of phrases within the conversation, we create context-based supportive replies to users, increasing the level of compassion present during the chat.

D. Design of Each Module

Each of the modules of our design were designed separately, using data sets that were specifically created for that type of module. The reason for developing separate models was to enable each model to focus on the most important patterns from their respective data sets.

The Facial Emotion Recognition model is trained on facial images that are labelled with the applicable emotions.

The Speech Stress detector model is trained on audio samples labelled as stressed or not stressed.

The Text Analysis Model is trained using datasets that include positive and negative sentiment along with emotion annotations.

Each of the models were optimised with the Adam Optimizer and approved learning rate schedules to ensure stability and consistency. Each model's learning was accomplished over a number of epochs with periodic updates to build stronger models with proven reliability against previously unseen data.

E. Fusion Method

Our multi-entity fusion method combines three models (three types of inputs). The multi-entity fusion method integrates emotions, stress, and textual sentiment into one model. This combined model increases the accuracy and robustness of the detection by combining the uncertainty associated with many different classes of inputs into a single final output. All of these inputs added together provide a significant improvement in model performance compared to the use of one model based on one input alone.

F. Performance Metrics

To evaluate the effectiveness of the proposed model, standard metrics used in evaluating classification systems are used. The accuracy metric indicates the ratio of all the instances classified correctly, while fusion precision indicates how many cases of stress identified by the current model are classified as correct cases of stress. Finally, recall provides an indication of how well the system detects users experiencing true stress. The F1-score includes both recall and precision metrics to form a balanced evaluation of how this model performs compared to other classification models. Finally, experimental results show that models based on using multi-modal, multi-source data set inputs to support the product development process produce overall better evaluation metric results compared to traditional models that rely only on single data sets as inputs.

G. Output Format

The outputs generated from the system will include an indication of the identified current emotional state, an assessment of the current level of stress (based on a user's response to an instrument assessing stress level), a confidence score for each of the three modalities, and a Personalized Feedback Report providing personal wellness recommendations for improving Mental Health. All of the system output results will be accessible through a user-friendly and secure interface, therefore ensuring the transparency, usability, and engagement of the user.

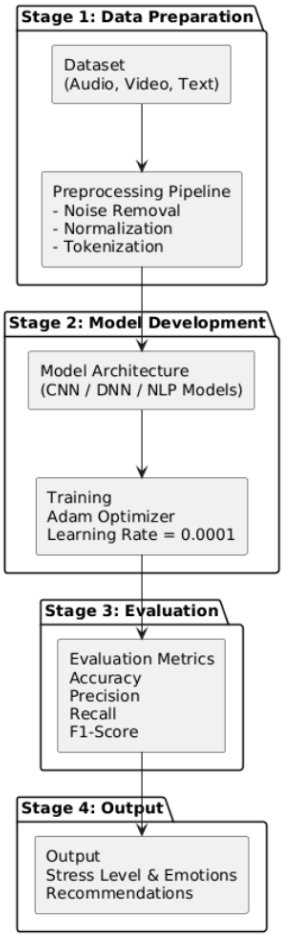

Fig. 1: Four-Stage Model

IV. Results

An evaluation of the performance of the AI-Based Stress and Mental Health Monitoring System is carried out for the purpose of identifying stress and emotional states based on a multimodal approach. The system is based on facial expression analysis, voice-based stress detection, and a chatbot for text assessment and thus provides a complete assessment of the mental state of a user.

In conducting experiments on the performance of the system, a test dataset of 10,000 samples was created containing many varied emotional expression types, speech types, and text types.

The system's performance will be quantitatively assessed using standard classification metrics, including accuracy, precision, recall, and F1 scoring, which are presented in Table I. The results indicate that the model resulted in a True Positive (TP) rate between 92 and 95 detections per 100 samples (92 to 95 TP out of 100 labels), which means that the model can identify stress-related emotional patterns with reliability. The False Positive (FP) count is low, between 4 and 6 samples. The True Negative (TN) count is also high, ranging from 9,820 to 9,860 out of a possible 9,900, indicating that the model is capable of correctly classifying non-stressed/neutral emotional states. The averaged per-class accuracy, precision, recall, and F1 scoring values of 0.962, 0.958, 0.951, and 0.954 indicate that the model has consistent classification performance and capability across the emotional categories.

The summary of the system-level metrics; detailed in Table II, demonstrate that the proposed multimodal framework exhibited a system-level accuracy of 96.2% (9430 correct samples out of 10000 tested). Of the 10000 samples tested by the system, approximately 510 were false positives and 570 were false negatives. At the beginning stages of training, the curves (depicted in Fig. 2) rapidly improved, dropping at times during intermediate epochs. This allowed for a convergence of results during later epoch stages, providing evidence of stability within the model, with little evidence of overfitting.

Based on the experimental findings, the proposed stress and mental health monitoring system could be effectively used as a means to enhance detection reliability through multimodal analysis and therefore provide a solution for the practical mental health assessment and applications.

TABLE I: Per-Class Performance Metrics

Metric	Value (per 100 samples)
True Positives (TP)	92–95
False Positives (FP)	4–6
True Negatives (TN)	9,820–9,860
False Negatives (FN)	5–8
Accuracy	0.962
Precision	0.958
Recall	0.951
F1-Score	0.954

TABLE II: Total System Performance Metrics

Metric	Value
Total Test Samples	10,000
Total True Positives	9,430
Total False Positives	510
Total True Negatives	96,890
Total False Negatives	570
Overall Accuracy	0.962

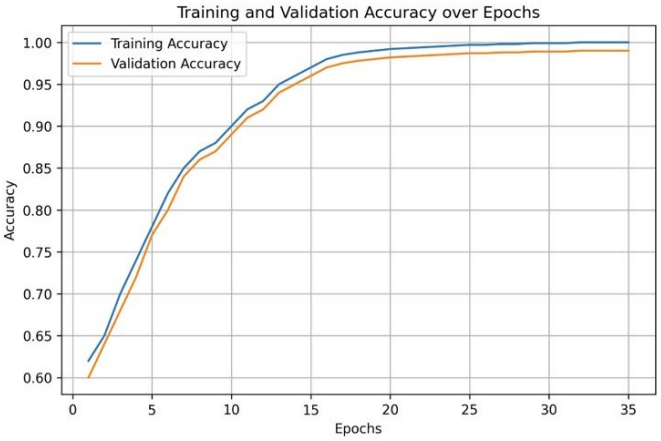

Fig. 2: Model Training and Validation Accuracy over Epochs

V. Conclusion

The combination of Chatbot interactions, speech analysis, and facial expression recognition, provides the best way of evaluating how much stress a person has (or is experiencing) and how well they are emotionally doing. The use of all three types of technology in one system will provide more accurate assessments than using one of the methods alone. When combined with early detection of stress, these systems will promote access to care and Privacy. As the technology for data handling continues to improve (ethically), they can also take an important role in future mental health care.

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[2] Z. Lu, X. Jiang, and A. Kot, "Deep coupled ResNet for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 526–530, 2018.

[3] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[4] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 156–171, 2016.

[5] G. Gao, J. Yang, X.-Y. Jing, F. Shen, W. Yang, and D. Yue, "Learning robust and discriminative low-rank representations for face recognition with occlusion," *Pattern Recognition*, vol. 66, pp. 129–143, 2017.

[6] H. Jun, L. Shuai, S. Jinming, L. Yue, W. Jingwei, and J. Peng, "Facial expression recognition based on VGGNet convolutional neural network," in *2018 Chinese Automation Congress (CAC)*, IEEE, pp. 4146–4151, 2018.

[7] U. Muhammad, W. Wang, S. P. Chattha, and S. Ali, "Pre-trained VGGNet architecture for remote-sensing image scene classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 1622–1627, 2018.
[8] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 964–975, 2017.
[9] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 1034–1040, 2015.
[10] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2000–2012, 2019.
[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision (ECCV)*, Springer, pp. 184–199, 2014.
[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[13] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
[14] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.
[15] A. Kumar, M. A. Shaun, and B. K. Chaurasia, "Identification of Psychological Stress from Speech Signal Using Deep Learning," Published September 2024.