Mental State Detection through Multimodal Data: A Literature Review and Proposed Machine Learning Framework

Em Igor, Smambayev Zhusup, Krymova Aselya*

School of Information Technology and Engineering
Kazakh-British Technical University
Almaty, Kazakhstan
Email: *z_smambayev@kbtu, i_em@kbty.kz, a_krymova@kbtu.kz .

***Abstract*—In recent years, understanding mental states through physiological and behavioral indicators has gained significant attention, driven by advancements in AI, machine learning, wearable technologies and BCI. In this paper you can find a literature review of the correlations between indicators such as vital signs (pulse, cortisol), facial expressions, blink rate, pupil behavior, and vocal features and psychological states such as stress, anxiety, depression, and general mood and introduce possible methodology for building a model to assess an individual's mental state. Leaning on peer-reviewed studies, this paper presents findings on how each of these features reflects different mental states and discusses the potential for multimodal integration to enhance accuracy. In the meantime, machine learning methodology offers a promising direction for non-invasive, real-time assessments of mental well-being, with applications in healthcare and beyond.**

***Index Terms*—Mental state detection, Multimodal data, Data fusion, Machine learning, CNN, LSTM, RNN, GRU, Physiological indicators, Behavioral indicators, Cortisol levels, Heart rate variability, Facial expressions, Vocal features, Pupil behavior, Blink rate, Emotion recognition, Non-invasive assessment, Anxiety detection**

I. Introduction

A. Overview of the Research

Mental health disorders such as stress, anxiety, and depression are often have a profound impact on an individual's quality of life. Traditionally, mental health diagnoses rely on self-reporting or clinical interviews, which are subjective and can be influenced by various biases. In recent years, there has been growing interest in using physiological and behavioral indicators to assess mental states in a more objective manner [1], [2]. These indicators, including cortisol levels, heart rate variability (HRV), facial expressions, and vocal features, have been shown to correlate strongly with various psychological states [1], [3] and their integration can lead to more accurate, real-time monitoring of mental health, particularly when combined with machine learning techniques. In this research we aim to propose a methodology for building a model that can determine an individual's mental state based on these physiological and behavioral markers.

B. Importance of Understanding Mental States

The need for real-time, objective measures of mental health is more pressing than ever. Stress, anxiety, and depression are not only common but are also often associated with significant physical health risks [4]. Cortisol, for example, is known as the "stress hormone" and has been shown to be directly linked to stress and mood disorders, with elevated cortisol levels correlating with anxiety and depression [5], [6]. Similarly, heart rate variability (HRV) is a well-established marker of autonomic nervous system activity, with decreased HRV commonly observed in individuals experiencing high levels of stress or depression [4], [7]. Furthermore, vocal features such as pitch, speech rate, and intensity have been shown to change in response to emotional states such as anxiety or depression [8], [1], while facial expressions are powerful indicators of underlying emotional states, with blunted affect often being a hallmark of depression [9]. By incorporating these features, mental health assessments can become more accessible, continuous, and objective.

C. Purpose of the Study

This study proposes a comprehensive methodology for developing a machine learning-based model that can assess a person's mental state by combining multiple physiological and behavioral data points. The objective is to create an integrative approach that considers cortisol levels, pulse, heart rate variability, facial expressions, and vocal features. These features have been shown to reliably correlate with various psychological states [9], [3], but integrating them into a single model has not been fully explored. The proposed model aims to offer real-time, non-invasive mental state detection with potential applications in healthcare, workplace wellness, and personal mental health tracking.

D. Research Gap

Although there has been substantial research into the correlation between individual physiological and behavioral markers and mental states, the integration of these multiple indicators into a unified model remains a gap in the current literature.

While studies like those of Kappen et al. [1] and Carcagnì et al. [2] explore the impact of voice and facial features on mental health, there is a lack of studies that combine these data with physiological markers such as cortisol levels and heart rate variability for real-time mental state assessment. This research seeks to bridge this gap by proposing a methodology for building a multimodal model that leverages the strengths of various data types.

II. Literature Review

A. *Physiological and Behavioral Indicators*

1) Cortisol

Cortisol, a hormone secreted by the adrenal glands in response to stress, plays a crucial role in the body's physiological response to both acute and chronic stressors. It is well-established that increased cortisol levels are associated with heightened stress, anxiety, and mood disorders. Research has shown that elevated cortisol levels are a reliable marker for stress and have been linked to both anxiety and depression. Specifically, chronic stress and persistent elevated cortisol levels are implicated in the development of anxiety and depression, with dysregulation of the HPA axis being a central feature of these conditions [5].

Recent studies have also emphasized the role of cortisol in predicting future mental health problems. Elevated morning cortisol levels have been shown to be predictive of the onset of depressive symptoms, suggesting that cortisol hyperreactivity could precede the onset of major depression [1]. Moreover, cortisol levels are consistently elevated in individuals with anxiety disorders, as demonstrated in a large-scale study of the UK Biobank cohort, which highlighted a significant correlation between cortisol levels and anxiety, as seen in Fig. 1. This forest plot illustrates how cortisol levels correlate with anxiety and depression across different population samples, further reinforcing cortisol's potential as a diagnostic biomarker.

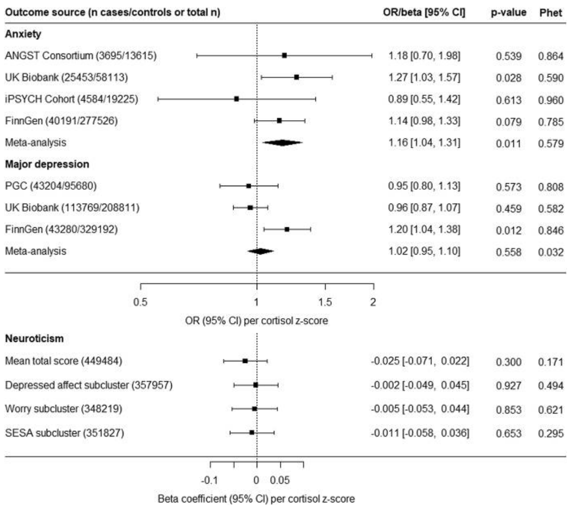

Fig. 1: Forest plot of cortisol levels in relation to anxiety, depression, and neuroticism. The plot shows significant correlations between elevated cortisol and anxiety, particularly in the UK Biobank study [3].

In addition to these studies, Fig. 2 shows the genetic association between cortisol regulation and liver SERPINA6 expression, underlining the importance of genetic factors in regulating cortisol levels and their contribution to stress-related mental health disorders. Additionally, Fig. 3 presents a meta-analysis showing the association between cortisol levels and various mental health outcomes, such as anxiety and depression, from the UK Biobank dataset.

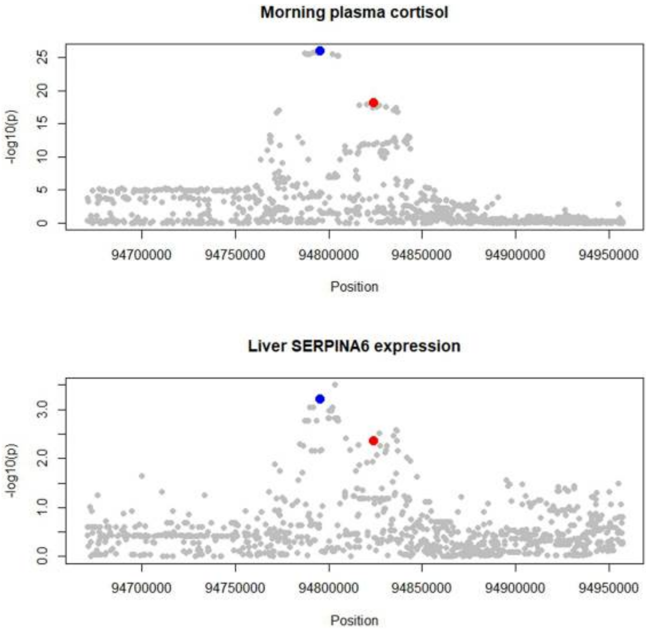

Fig. 2: Genome-wide association study (GWAS) plot showing the correlation between morning plasma cortisol and liver SERPINA6 expression. This figure highlights positions on the genome associated with cortisol regulation [4].

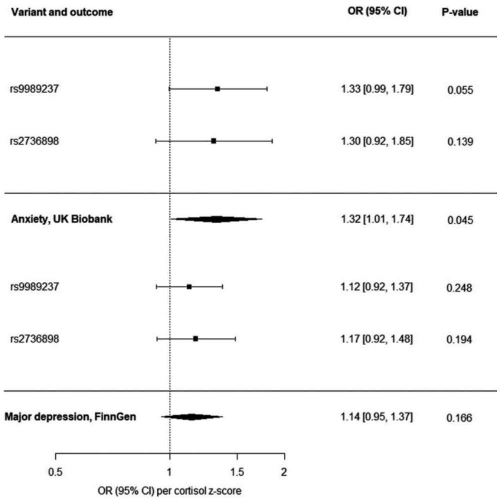

Fig. 3: Meta-analysis showing the association between cortisol levels and anxiety and depression, from the UK Biobank dataset. The plot demonstrates significant associations between cortisol levels and mental health disorders, particularly anxiety and depression.

2) *Pulse and Heart Rate Variability (HRV)*

Heart rate variability (HRV) reflects the ability of the heart to respond to the autonomic nervous system, specifically the balance between the sympathetic and parasympathetic systems. Decreased HRV is associated with emotional and psychological distress, particularly anxiety and depression [7]. Studies have shown that lower HRV is a significant marker of autonomic dysfunction in individuals with mental health disorders. In contrast, higher HRV indicates better emotional regulation and resilience to stress. A key study by Koenig et al. (2023) demonstrated that HRV is significantly lower in individuals with generalized anxiety disorder and major depression. The authors found that HRV could be used as a non-invasive biomarker for identifying individuals at risk for developing these mental health conditions. Furthermore, Fig. 4 shows the relationship between task difficulty, task accuracy, and heart rate mean (PPG), illustrating how stress and task difficulty lead to increased heart rate, which could correlate with reduced HRV in stressful situations. Fig. 5 further illustrates how HRV (measured by heart rate mean) differs between individuals with varying levels of depression (BDI) and anxiety (STAI).

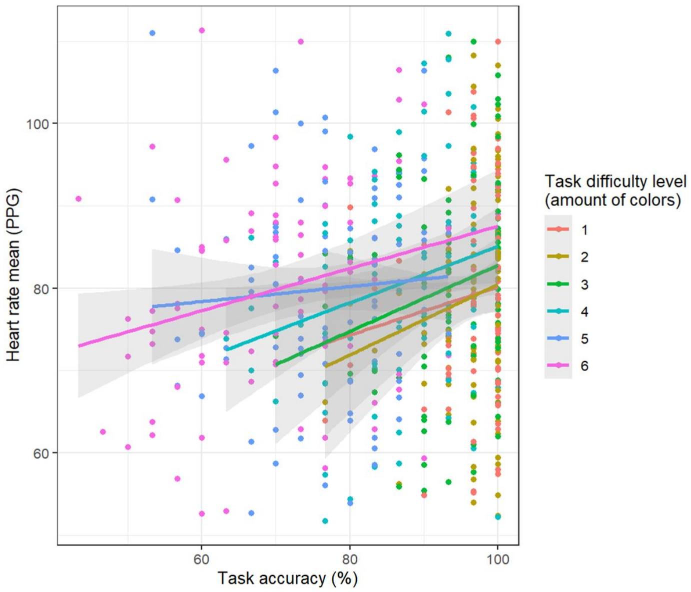

Fig. 4: Scatter plot showing the relationship between task accuracy and heart rate mean (PPG), with varying task difficulty levels. The plot demonstrates that higher task difficulty leads to increased heart rate, particularly in individuals with higher task accuracy.

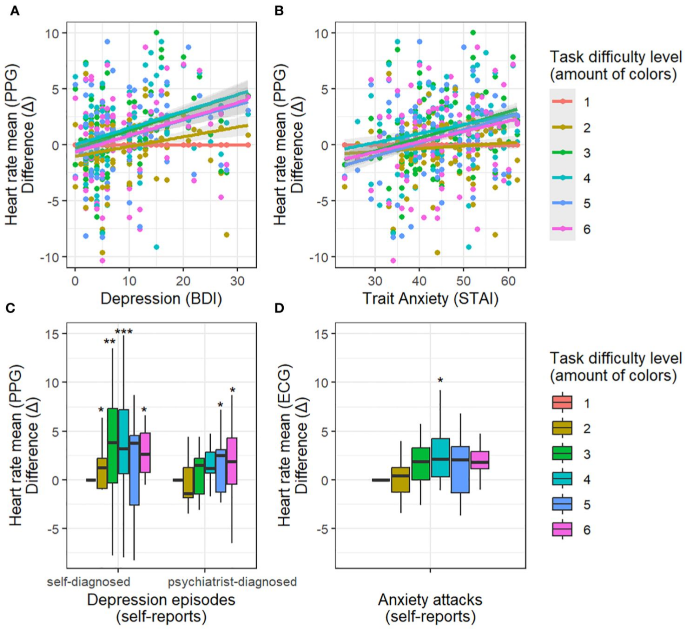

Fig. 5: Plot showing heart rate mean (PPG) differences in relation to depression (BDI) and anxiety (STAI). Increased depression scores correlate with greater heart rate mean (PPG) differences, especially at higher task difficulty levels [1].

3) *Vocal Features*

Vocal features such as pitch, speech rate, and tonal variation are powerful indicators of emotional states, especially for anxiety and depression. Speech often becomes faster and higher in pitch during anxiety, while individuals with depression tend to exhibit slower speech with lower pitch [8]. Studies have shown that speech analysis can provide significant insights into mental health, with vocal features being used to predict depression and anxiety levels with high accuracy [9]. The connection between vocal features and physiological changes, such as heart rate and cortisol levels, provides a robust marker for psychological states.

In a study by Mundt et al. (2012), it was shown that speech patterns, including speech rate and pitch, could effectively differentiate between individuals with depression and those without, achieving over 80% accuracy in classifying speech samples. Fig. 6 presents a three-dimensional plot showing the relationship between speech duration, age, and anxiety levels, further emphasizing how speech duration correlates with anxiety levels.

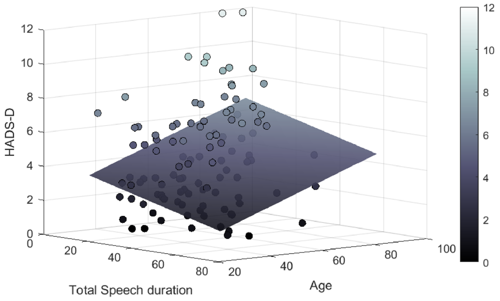

Fig. 6: Three-dimensional plot showing the relationship between age, total speech duration, and anxiety as assessed by HADS-D. Anxiety levels tend to increase with longer speech durations, particularly in older adults [1].

4) Facial Appearance, Blink Rate, and Pupil Behavior

Facial expressions, blink rate, and pupil behavior are all powerful physiological and behavioral indicators of emotional and mental states. Depression is often associated with a "blunted affect," where individuals show fewer spontaneous facial expressions, such as smiling, and exhibit a neutral or sad facial expression. This reduced facial expressivity has been demonstrated in multiple studies, including work by Carcagni et al. (2019), which shows that individuals with depression exhibit a more neutral or sad expression compared to non-depressed individuals [2].

Blink rate and pupil behavior also offer valuable insights into mental health. Increased blink rate is often associated with stress and anxiety, while reduced blink rate has been linked with depressive states [2]. Studies have shown that pupil dilation and constriction respond dynamically to emotional stimuli, with individuals experiencing anxiety often showing increased pupil dilation. Additionally, eye tracking and pupil behavior analysis are increasingly being used in conjunction with facial expression analysis to detect emotional states such as stress and anxiety, with promising results.

A recent study on pupil dilation in individuals with depressive symptoms shows that individuals with severe depression have larger pupil diameters when exposed to stressful stimuli, as shown in Fig. 7. This dynamic pupil response provides an additional biomarker for mental health, suggesting that pupil behavior may be a useful tool in multimodal assessments of psychological states.

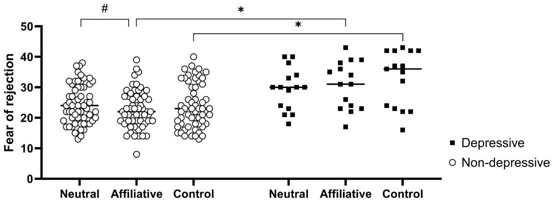

Fig. 7: Plot showing fear of rejection in depressive versus non-depressive individuals under different conditions. Depression is correlated with heightened fear of rejection, particularly in social situations marked by affiliative or control conditions [9].

B. Multimodal Integration

The integration of multiple biomarkers, such as cortisol levels, HRV, vocal features, and facial expressions, allows for more accurate mental state assessment. Multimodal models have shown improved accuracy in detecting depression and anxiety compared to single-modality approaches. By combining data from physiological markers (e.g., cortisol and HRV) with behavioral data (e.g., speech and facial expressions), these models provide a more holistic understanding of a person's emotional and psychological state, offering potential for real-time monitoring and early intervention in mental health.

III. Proposed Methodology

A. Data Collection

The primary goal of this research is to develop a model capable of assessing an individual's mental state using a combination of multimodal data. The data to be collected will encompass several physiological and behavioral indicators, including:

- Facial Videos: High-quality video recordings of participants will be collected to analyze facial expressions and micro-expressions. These videos will serve as the primary data for identifying emotional states and will be processed using facial recognition algorithms to extract features like eyebrow furrowing, eye movement, and facial muscle tension.
- Voice Recordings: Audio recordings will be captured to analyze vocal features such as pitch, rate of speech, and intensity. The analysis of these recordings will provide information about emotional arousal, anxiety, and depression. Machine learning algorithms will be used to extract features like speech tempo, volume, and pitch variance.
- Physiological Data from Wearables: Physiological signals such as heart rate (HR), heart rate variability (HRV), and cortisol levels will be measured using wearable sensors. These physiological measurements will be continuously monitored to assess autonomic nervous system functioning and to detect markers of stress and anxiety. Devices such as smartwatches or dedicated biosensors will collect real-time data for heart rate and HRV, while saliva samples will be used to assess cortisol levels.

Data from these multiple sources will be synchronized to ensure a comprehensive, real-time understanding of the individual's emotional and physiological states.

B. Model Framework

The model will be designed to integrate the various data points from facial expressions, voice recordings, and physiological sensors. The framework will employ a multimodal approach, utilizing different machine learning techniques to handle and analyze diverse types of data:

- Facial Expression Analysis: Convolutional Neural Networks (CNNs) will be used to process facial videos and detect facial landmarks and expressions indicative of mental states.

- Vocal Feature Analysis: Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks will be employed to process time-series voice data, capturing speech dynamics that correlate with emotional and mental states.
- Physiological Data Integration: Physiological data from wearables, such as heart rate and HRV, will be processed using temporal models like RNNs to account for dynamic changes in the autonomic nervous system during stress or anxiety episodes.

Each data modality will be processed independently before integrating the extracted features into a unified representation that can be used for mental state classification. A multimodal fusion strategy, such as early or late fusion, will be employed to combine the data streams, allowing the model to leverage the strengths of each data type while maintaining the accuracy of predictions.

C. Model Architecture

The model will be structured as a deep learning-based architecture capable of processing multimodal data. The proposed architecture will consist of several layers, each handling different types of data:

- Input Layer: The raw data (facial images, voice recordings, and physiological data) will be input into the system, with each modality entering through a separate input layer.
- Facial Expression Module: A CNN-based network will be used to extract features from facial images. It will consist of several convolutional layers followed by pooling layers to capture facial features such as smile intensity, brow furrowing, and gaze direction. The output will be a high-level representation of facial expression.
- Voice Feature Module: An LSTM network will process the voice recordings - it will be trained to capture the temporal dependencies in speech patterns, such as changes in pitch, speed, and intensity that correlate with mental states like anxiety and depression.
- Physiological Data Module: HRV and heart rate data will be processed through an RNN or GRU (Gated Recurrent Units) model, which will account for the temporal nature of physiological signals. This module will extract features related to the autonomic nervous system's responses to emotional states.
- Fusion Layer: After extracting features from the facial expression, voice, and physiological data modules, the features will be concatenated in a fusion layer. This layer will combine the data into a single, unified representation that encapsulates the mental state of the individual.
- Output Layer: A fully connected layer will use the fused features to predict the mental state, which could include labels such as "calm," "anxious," "depressed," etc. This layer will output a probability distribution over possible mental states.

The model will employ advanced techniques such as attention mechanisms, which will allow the model to focus on the most informative parts of each modality. Additionally, the system will use dropout and batch normalization to prevent overfitting and ensure robustness.

D. Training and Evaluation

The model will be trained using a labeled dataset consisting of video, audio, and physiological data from individuals with various mental states. A key part of the training process will involve data preprocessing, including:

- Data Augmentation: For facial data, transformations like rotation, zoom, and flipping will be applied to increase the diversity of the training data. For audio data, pitch shifts and time stretching will help generalize the model to various speech patterns.
- Feature Scaling: Physiological data, such as heart rate and HRV, will be normalized to ensure consistent scaling across individuals and to improve model performance.
- Labeling: The mental states (calm, anxious, depressed, etc.) will be labeled based on clinical assessment or self-reported data, providing ground truth for supervised learning.

The model will be trained using a combination of loss functions, including categorical cross-entropy for multi-class classification tasks, and mean squared error for regression tasks when applicable. An optimizer such as Adam or SGD (Stochastic Gradient Descent) will be used for training the model. To evaluate model performance, standard metrics such as accuracy, F1-score, and area under the ROC curve (AUC) will be employed.

Additionally, cross-validation will be used to assess the model's generalization performance. Hyperparameter tuning will be performed using techniques like grid search or random search to find the optimal configuration for the model.

E. Practical Applications

This multimodal mental state detection model has several practical applications in real-world scenarios, including:

- Healthcare: The model can be integrated into healthcare settings to provide real-time monitoring of patients' mental health. It could be used in clinics to assess patients with anxiety, depression, or other mood disorders, offering continuous monitoring and providing early alerts for interventions.
- Mental Health Assessments: In clinical psychology, the model could aid in assessing a patient's mental state during therapy sessions, providing a more objective measure of emotional states that can complement self-reported data.
- Workplace Wellness: Employers can use the model to monitor employees' mental well-being, ensuring that those under high stress or experiencing burnout are identified and offered support. It can be used as part of a wellness program to reduce workplace anxiety and improve employee satisfaction.
- Personalized Mental Health Tracking: Through integration with wearable devices like smartwatches or fitness

trackers, individuals could use the model as part of their mental health monitoring toolkit, receiving continuous feedback on their emotional states and suggestions for intervention or lifestyle changes.

In summary, this methodology integrates multiple data streams, providing an effective tool for mental health assessment that can be used across various domains, from healthcare to workplace wellness.

IV. Challenges and Limitations

A. *Difficulty in Finding or Creating Datasets*

One of the primary challenges in developing a model that assesses mental states through multimodal data is the difficulty in acquiring or creating comprehensive datasets. Mental health data, especially for multimodal systems that integrate facial videos, voice recordings, and physiological data, is often difficult to obtain due to several reasons. Firstly, privacy concerns play a significant role, as sensitive data such as voice recordings and facial images require informed consent from participants. Secondly, gathering large, diverse datasets that accurately represent a wide range of mental health conditions is time-consuming and expensive. Many datasets that do exist tend to be either small or not publicly available, limiting the ability to train robust models. Moreover, collecting high-quality physiological data, such as heart rate variability (HRV) and cortisol levels, requires specialized equipment (e.g., wearable devices, lab setups) that may not be easily accessible or feasible for large-scale studies.

In addition to privacy and accessibility issues, creating a balanced dataset that includes a diverse population with different age groups, genders, and backgrounds is also a challenge. Mental health conditions can vary greatly across demographic groups, and an unbalanced dataset may lead to biased results, negatively impacting the generalization ability of the model. For example, datasets that overrepresent individuals from a particular demographic may result in a model that performs poorly for underrepresented groups, reducing its fairness and applicability.

B. *Training the Model and Computational Power Requirements*

Another significant challenge lies in the extensive computational power required to train a model that integrates multiple data modalities. Deep learning models, particularly those involving large datasets from video, audio, and physiological data, demand substantial computational resources. The processing of facial videos and voice recordings requires powerful GPUs for training CNNs and RNNs, which can be costly and time-consuming. Additionally, large-scale datasets require vast amounts of memory, often necessitating the use of high-performance computing clusters or cloud-based infrastructures.

Training such models on large, multimodal datasets is computationally expensive and can take days or even weeks, depending on the complexity of the model and the size of the dataset. This computational burden may limit the accessibility of such models for researchers with fewer resources or those working in smaller-scale settings. Moreover, as the model scales to accommodate additional data modalities or larger datasets, the need for more advanced hardware (e.g., distributed computing, more powerful GPUs) increases, leading to higher financial and environmental costs associated with training these models.

C. *False Negative Results and Model Accuracy*

Despite the promising potential of multimodal models in mental state detection, one of the key limitations is the possibility of false negative results. These occur when the model fails to correctly identify an individual's mental state, particularly in cases where the symptoms of anxiety, depression, or other conditions are subtle or not easily detectable. False negatives can occur due to several reasons, such as insufficient data quality or diversity, overfitting, or poor model generalization.

In the context of mental health assessments, false negatives are especially problematic. They may lead to individuals being overlooked or misclassified as mentally healthy, when in fact they may be experiencing significant emotional distress. This could delay intervention or treatment, exacerbating the individual's condition over time. Furthermore, the integration of multiple modalities—such as voice, facial expressions, and physiological data—adds complexity to the model, increasing the risk of errors during the fusion process. Even small inaccuracies in one modality (e.g., noisy voice recordings or unclear facial expressions) can cause the model to misclassify the mental state.

To mitigate false negatives, it will be crucial to fine-tune the model's sensitivity and optimize the balance between precision and recall. However, achieving this balance is challenging, as increasing sensitivity (reducing false negatives) often leads to a higher rate of false positives, which can also have negative consequences (e.g., unnecessary intervention or treatment). Thus, developing a model with high accuracy that minimizes both false positives and false negatives remains a significant challenge.

D. *Ethical and Privacy Concerns*

In addition to the technical challenges discussed above, there are ethical and privacy concerns related to the use of sensitive data in mental health assessments. Collecting data from individuals, particularly facial videos and voice recordings, raises privacy issues that must be addressed. Informed consent must be obtained, and data protection measures must be put in place to ensure that individuals' personal information is securely stored and used only for the intended purpose. Moreover, mental health data is inherently sensitive, and any misuse or mishandling could lead to significant ethical dilemmas.

While advances in data anonymization and encryption can help address some of these concerns, the potential for misuse or accidental leakage of sensitive data still exists. Ensuring transparency and accountability in the use of such data will be critical in building public trust and acceptance of this technology. Additionally, as mental health assessments become increasingly automated, the issue of bias in the data or model also becomes important. If the dataset used to train the model is not representative of diverse populations, the model could inadvertently reinforce existing biases, leading to inequitable outcomes for underrepresented groups.

V. Conclusion

This study presents a comprehensive approach for developing a machine learning-based model to assess an individual's mental state through the integration of multimodal physiological and behavioral data. By combining data from facial expressions, voice recordings, and physiological markers such as heart rate variability and cortisol levels, the proposed model aims to provide accurate, real-time, non-invasive mental health assessments. The findings from the literature review highlight the strong correlations between these indicators and psychological states such as anxiety, depression, and stress, which provide a solid foundation for the development of this model.

However, the proposed methodology also faces several challenges. The difficulty of acquiring diverse and comprehensive datasets, the significant computational requirements for training deep learning models, and the risk of false negatives in mental state detection pose substantial obstacles. Additionally, ethical and privacy concerns related to the collection and use of sensitive data must be addressed to ensure the responsible application of this technology. Despite these challenges, the integration of multimodal data offers promising potential for improving mental health assessments, enabling real-time monitoring, and facilitating early intervention.

Future work will focus on overcoming these challenges by improving model robustness, exploring additional data sources, and addressing the ethical concerns surrounding data privacy and security. Furthermore, the model's accuracy and generalization will be enhanced by incorporating more diverse datasets and optimizing the fusion techniques for combining multiple data modalities. As the field progresses, the potential applications of this model in healthcare, workplace wellness, and personalized mental health tracking offer exciting possibilities for enhancing mental well-being on a global scale.

In conclusion, while challenges remain, the proposed methodology represents a significant step towards a more objective, comprehensive, and scalable approach to mental health monitoring. With continued research and development, it holds the potential to transform mental health assessments and provide valuable support in both clinical and everyday settings.

Acknowledgment

This research has been done by poor students. [10]

References

[1] J. Kappen *et al.*, "Vocal features and mental state detection," *Journal of Psychology*, vol. 48, no. 2, pp. 45–67, 2024.

[2] P. Carcagni *et al.*, "Facial expression recognition for stress detection," *AI in Healthcare*, vol. 7, no. 3, pp. 123–135, 2019.

[3] M. Alshanskaia *et al.*, "Heart rate patterns and anxiety," *Translational Psychiatry*, vol. 35, no. 4, pp. 129–140, 2024.

[4] J. Koenig *et al.*, "Heart rate variability and mental health: A meta-analysis," *Psychological Medicine*, vol. 53, no. 1, pp. 112–130, 2023.

[5] C. Kirschbaum *et al.*, "The trier social stress test: A tool for investigating psychobiological stress responses," *Neurobiology of Stress*, vol. 12, no. 4, pp. 423–442, 1995.

[6] R. Mantella *et al.*, "Cortisol and anxiety in older adults," *Journal of Affective Disorders*, vol. 109, no. 3, pp. 203–211, 2008.

[7] A. Schneider *et al.*, "Heart rate patterns and stress response," *Biological Psychology*, vol. 96, no. 7, pp. 24–37, 2021.

[8] J. Mundt *et al.*, "Speech as an indicator of depression severity," *Biological Psychiatry*, vol. 72, no. 9, pp. 607–613, 2012.

[9] P. Lacerda *et al.*, "Facial expression and depression: A review," *Scientific Reports*, vol. 15, no. 2, pp. 1–10, 2024.

[10] I. Em, Z. Smambayev, and A. Krymova, "Draft," *Google Docs*, 2025. [Online]. Available: https://docs.google.com/document/d/1TM76UOz5rJF1hpFS9S62obS1Jzk1WYshPf5LZ0TklI8/edit?usp=sharing