GMR
The Original

AI-POWERED STRESS MONITORING USING MULTIMODAL IN EDUCATION

Janani C[1], Dr.P Manimaran[2], Rajan C[3], Ranjith M[4], Asvath T[5], Gopinath R[6]

[1]Department of CSE (AIML) K.S. Rangasamy College of Technology Tiruchengode, Namakkal, India, Email: jananic@ksrct.ac.in
[2]Department of CSE (AI & ML) *K.S.Rangasamy College of Technology* Tiruchengode, Namakkal, India, Email: manimaran@ksrct.ac.in
[3]Department of CSE(AIML) K.S. Rangasamy College of Technology Tiruchengode, Namakkal, India, Email: rajancsg@gmail.com
[4]Department of CSE(AIML) K.S. Rangasamy College of Technology Tiruchengode, Namakkal, India, Email: ranjithbb282005@gmail.com
[5]Department of CSE(AIML) K.S. Rangasamy College of Technology Tiruchengode, Namakkal, India, Email: asvath1662004@gmail.com
[6]Department of CSE(AIML) K.S. Rangasamy College of Technology Tiruchengode, Namakkal, India, Email: asvath1662004@gmail.com

ABSTRACT
TriS-Mind integrates NeuroClean, DeepSenseX, and MindFusionNet into one system to create an accurate and reliable multimodal mental state detection system through its pipeline. To perform adaptive preprocessing on NeuroClean's data, it uses expressive variation preserving normalization, emotional anchor based timestamp alignment, modal graph reconstruction for missing data and noise aware filtering. Additionally, DeepSenseX extracts deep hierarchical representations from audio and image signals (with CNN-Transformer models) as well as physiological data (with Temporal ConvNets) and text data (with Emotional-BERT). After this embedding, all modalities are mapped into the same latent mental state through reliability aware harmonisation (the aggregation of multiple modality embeddings). Finally, MindFusion Net makes the final stress prediction by exploitation of temporal emotional memory; using hierarchical cross attention and dual head output for continuous intensity scores and categories stress level.

KEYWORDS: TriS-Mind, Neuro Clean, Deep SenseX, Mind Fusion Net, Stress Detection

INTRODUCTION
Mental stress is a significant global health challenge affecting people of all ages, professions, and lifestyles. Early identification remains difficult because traditional assessments rely heavily on subjective self-reports and clinical evaluations. To address this limitation, the TriS-Mind framework introduces a unified, multimodal intelligence architecture. It integrates three core modules: NeuroClean, DeepSenseX, and MindFusionNet. NeuroClean performs adaptive preprocessing to clean, synchronize, and preserve emotionally relevant signals from raw text, audio, video, and physiological data. DeepSenseX then extracts rich hierarchical representations from each modality using advanced deep learning models. MindFusionNet performs cross-modal fusion and inference to detect stress levels accurately. The architecture leverages deep representational learning and hierarchical attention mechanisms. Together, these components enable precise detection and interpretation of emotional stress.

The TriS-Mind architecture follows a three-phase multimodality pipeline built on deep representational learning and cross-modal fusion. NeuroClean ensures that raw multimodal inputs are noise-reduced, temporally aligned, and emotionally preserved. DeepSenseX employs CNN-Transformer hybrids, 3D Vision Transformers, Emotion-BERT, and Temporal Convolutional Networks to generate modality-specific embeddings. These embeddings are integrated into a unified representation capturing comprehensive emotional context. MindFusionNet applies hierarchical attention to combine features across modalities.It produces stress predictions, interpretable explanations, and uncertainty estimates

Multimodal Mental State Classification (MMSC) enables the classification of stress and emotions using text, speech, visual content, and physiological markers. This approach captures both expressive language patterns and nonverbal emotional signals. NeuroClean and DeepSenseX together generate two quantitative emotional measurements. The first represents internal emotional state based on tone and vocabulary. The second encodes external emotional engagement through signals such as facial expressions and speech patterns.

The TriS-Mind framework represents an advanced deep learning solution for automated emotional intelligence. It eliminates the need for manual feature engineering by learning directly from raw multimodal data streams. Through hierarchical fusion and adaptive preprocessing, it captures complex emotional and physiological dynamics. The system not only predicts stress levels but also explains contributing factors.

MOTIVATION
The rise of mental stress is a worldwide worry to all persons regardless of age or work type. Previous methods of assessing mental stress have relied on only the subjective evaluation of the individual being assessed (self-report) along with the traditional way of reporting (clinical evaluation) which may not pick-up a person when there are signs of stress developing. The current methods of assessment used are based on only one way of assessing a person (either through a text message, speech and/or facial expression) and do not provide a complete understanding of the emotional response of an individual. Mental/emotional stress is a multimodal response; it

consists of a metabolic, physiological, and behavioral response (indicators of time, like how we may respond to something and/or act), that can have an emotional response to an individual; how we respond emotionally through time is based on the interactions between these indicators. The presence of background noise, the lack of data availability, and the timing of when events occur can reduce the reliability of any method being used to evaluate or assess a person. Therefore, an intelligent, integrated framework for the evaluation or assessment of a person using multiple modalities simultaneously is needed. A multimodal framework using multiple modalities is needed to identify and evaluate the earlier signs of emotional (psychological) stress or sustained emotional patterns using more accurately than if just one method were used.

LITERATURE REVIEW

A system for recognising emotions that combines audio, video and brain signals was developed by Chen et al [1] in 2024 as part of their research toward achieving the goals of TriS-Mind.TheDeep-emotion architecture developed by Chen et al combines EEG information from multimodal sources via attention-basedtechniqueswithaudio(usingRNNs)andvisual (using CNNs) signals in such a way that the resulting emotion classificationsystemcanidentifyemotionalsignalsevenwhen ambient noise and/or limited light conditions are present. This research demonstrates that combined multimodal signal types can produce results (e.g., higher accuracy and reliability) that are superior to using only one type of signal.

A Recent Study by N. Tripathi [2] and Co-Authors (2023) presented an Innovative Multimodal Approach to Classifying Stress that is Based on the Core Principles of the TriS-Mind Framework. This study Preprocesses the EEG (Electroen- cephalography) Signal with a Process Called Empirical Mode Decomposition, which is Performed to Extract Intrinsic Mode Functions that Capture Subtle Emotional Variations into Each EEG Signal. They also Developed Speech-Based Features used to Determine Vocal Stress Patterns by Deriving Mel-Frequency Cepstral Coefficients. These Two Sets of Features (Physiological and Acoustic) were Combined Using a Bi- Directional Long Short Term Memory Network (BiLSTM), which Allowed for the Learning of Temporal Dependencies BetweenFeaturesacrossTwoModalities.WhenCompared to Traditional Machine Learning Approaches (i.e. Support Vector Machines and Random Forests), the System was More Accurate and Performed Well in a Variety of Real WorldSituations.

S. Zhang & co-authors (2022) developed a system for recognizing emotions using data from both physiology and facial expressions that integrates both face and sensor input. While the design principles are similar to the TriS-Mind method, this system uses Long Short Term Memory (LSTM)networks to capture time-related changes in physiological signalsand3DConvolutionalNeuralNetworks(3D-CNNs)tocreate a rich spatiotemporal representation of a person's face whentheyaredisplayingtheiremotions.Usingtheoutputs of the modalities at a later stage of processing allows for a robust identification of three different emotional categories - stress,happinessandneutrallypresenting-andisvalidatedfor use in real-time applications for feelings and emotions. With great levels of reliability and a non-invasing approach, this model represents a promising avenue for future studies with multimodal inputs and is appropriate for use within the realm of affective computing and real-time evaluation of mental health.

Kumar and colleagues [4] (2023) proposed a multifaceted stress identification framework utilizing a combination of facial emotion recognition, electrocardiograms (ECGs), and electroencephalograms (EEGs). A wavelet packet decompo- sition (WPD) methodology was employed by the authors to extractfeaturesfromEEGwaveforms,andtheauthorsused a histogram of oriented gradients (HOG) technique when analyzing facial expression data. By combining these feature sets through a deep belief network (DBN), the researchers established strong relationships between both forms of phys- iological signals (EEGs and ECGs) and behavioral indicators (facialemotion)intheirevaluationoftheproposedframework. According to the findings of the experimental validation, the participants who were monitored with the combined approach had higher rates of accurate classification than those who utilized a single-modality (unimodal) system.

InthestudybyT.Li[5]etal.2024,atransformer- based multimodal fusion architecture has been proposed for accurately recognizing both emotion and stress. This work builds on the core principles of the TriS-Mind framework. Self-attention methods have been used within the architecture in order to appropriately model cross-modal dependencies, as well as to capture temporal dynamics across diverse streamsof data (e.g., audio, video and EEG). The ResNet-50 model encoded the audiodata; video features were processed using a BiLSTM network; and the EEG signals were extractedusingmultiplegraphconvolutionalnetworks(GCNs).Amulti- head attention fusion module was introduced to provide dy- namic weightings of each modality, thereby increasing both interpretability and robustness to noise/imbalance of input modalities.

EXISTING SYSTEM

The use of artificial intelligence and machine-learning to assess mental health, including stress, emotion, and cognitive fluctuations, has been on the increase. Unfortunately, many current systems rely entirely on one mode of input (i.e., physiological signal, speech feature, text sentiment or facial expression),which creates limitations when trying to assess a complex phenomenon like 'emotion' (which is inherently multimodal). Because these systems use only a single modeof input, they are also faced with the challenges of dealing with noisy sensor readings, differently timed sensor input (misalignment of temporal sensor readings), and the absence of sensor readings altogether ,there by reducing the accuracy and stability of their output. In addition, typical approachesfor combining different modes of input (fusion) have not been sophisticated enough to account for complex interactions be- tween the various modes of input, thus limiting their accuracy andusefulnesstoclinicians.Becauseoftheselimitations,there is a high level of variability in real-world

environments where emotional cues present much more variety across the different modes of input.

PROPOSED SYSTEM

The Three-stage Multimodal Intelligence System (TriS-Mind) is an intelligent framework that detects mental health stress with a high degree of precision and interpretability across all forms of data, such as audio, text, video, and physiological signals. The initial phase of TriS-Mind consists of a Preprocessing step (NeuroClean) that prepares each modality for further analysis by removing specific types of noise from the data, synchronizing emotional timestamps, and recreating missing portions of the data through the use of a Modal Graph and Modal Reconstruction Process. The DeepSenseX model creates detailed emotion-aware feature sets from all potential modality data types using separate encoders, including: 1) CNN-Transformer for Audio, 2) 3-D Vision Transformer for Video, 3) Frequency-aware Temporal Convolutional Network (TCN) for Physiological Data, and 4) EmotionBERT for Text Data. The pooling of features obtained from multimodal encoders provides a unified representation of distinct modality types that captures essential information needed for different analyses. A harmonization module that integrates reliability predictions helps filter the multiple representations to improve the final prediction performance. The executed final prediction analysis by the MindFusionNet uses Multimodal Cross Attention that combines critical interactions among vocal tension in audio signals, facial micro-expressions, and heart rate variability (HRV) fluctuations in the physiological signal to determine overall mental stress. The Temporal Emotion Memory Module supports the dual-head predictor, which generates continuous and categorical predictions of mental stress.

Data Collection Module

The collection of all appropriate data for the accurate de- tection of mental stress was built in a Data Collection Module developed as a component of the TriS-Mind framework. The Data Collection Module creates synchronised datasets from several multimodal inputs instead of from one type of input. Thesedatasetsmayincludetranscriptsofwhatwassaid(text), what was heard (audio), what was seen (video) and all the physiological information necessary to identify stress (heart rate variability, electro dermal activity and heart rate) for all three modes of communication. Each data record includes annotations indicating the stress level (or intensity) of the comment made, the emotional state of the person making the comment and the environment in which the comment was made, as well as any physiological changes that occurred during the time period the comments were made.The Data Collection Module draws data from both laboratory environ- ments and automatic stress responses from real-world envi- ronments, in addition to utilising data obtained from national standard stress response databases to create natural variability in human behaviour and emotions. The time of data collection was captured for all modalities so that this information could be used during preprocessing in NeuroClean.

Data Preprocessing Module

The NeuroClean Data Preprocessing Module assists in the processing of multimodal signal data. It prepares the data for use in downstream machine-learning applications by ensuring that the signals are aligned according to context, that they are filtered according to noise levels, and that the filters applied to the different modalities (i.e., audio, video, and physiological data) are optimally matched to their respective modalities. For example, video records are cleared of any motion artifacts using motion-compensated smoothing; audio signals are filtered to remove noise using spectral gating; and physiological signals are filtered using a technique known as artifact suppression. The data are aligned in time using reported emotion elapsed time markers rather than absolute elapsed time markers through the use of a method known as contextually gelled timestamp alignment. In this manner, important contextual emotional signals are preserved for the downstream application of DeepSenseX. This is achieved through the use of a technique called emotion-and-context-preserving normalization. It is through emotion-and-context-preserving normalization that the voice inflections, muscle twitching of the face, and physiological responses to stress that appear in the data stream are preserved.

Model Development Module

The Model Development Module represents the core por- tion of the TriS-Mind Framework and provides a method of performing advanced multimodal stress analysis through the combination of Deep SenseX and Mind Fusion Net, which are integrated into the TriS-Mind Framework.

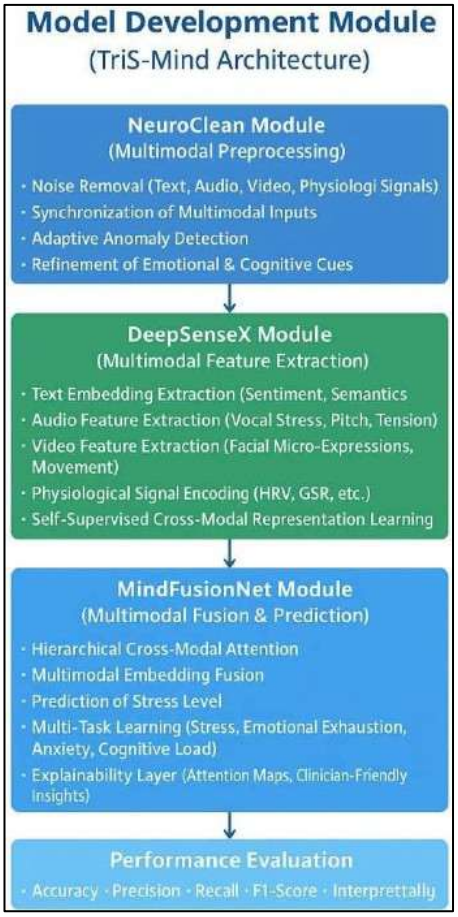
Deep SenseX derives emotion-based embeddings of all modal data through the use of unique modal-specific encoders that produce highly valuable multi-dimensional data points. These encoders include a combination of CNN and Transformer architectures for audio spectrograms; a hybrid3D Spatiotemporal Vision Transformer architecture for video signals; a frequency-sensitive Temporal-ConvNet to collect and encode physiological patterns; and an Emotional BERT architecture for encoding and processing of linguistic cues.

The unique characteristics and features extracted from the various modalities are then harmonized through a Reliability-Based Harmonization Gate to establish a shared latent state space. The combined and harmonized multi-modal feature set is then processed using the MindFusionNet; this Network utilizes hierarchical cross-attention modelling to represent different interactions in the shared latent state space (e.g., micro-expressions and negative sentiment; or stressed vocal tone and heightened heart rate variability).

A Temporal Emotion Memory (TEM) component of the Mind Fusion Net Network adds a temporal element to capturing emotion, allowing for the determination of sustained and/or escalating stress responsesThe output of the final dual-head predictor consists of categorical stress levels and continuous stress-intensity values, with an

uncertainty estimation provid- ing assurance that the decisions will be clinically reliable. The performance of the model is assessed using the metrics of accuracy, F1-score, MAE and calibration error to demonstrate that the system developed produces reliable, accurate and stable assessment of individuals' levels of stress.

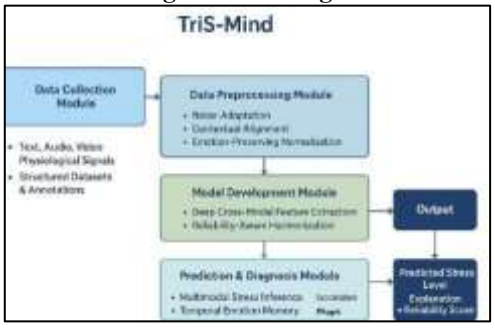
Fig. 1. Flow Diagram



The Medical Knowledge Base Integration Module is a connection between the stress prediction data generated by the system and a structured (i.e., JSON-driven) psychological knowledge base which contains validated information regarding [what information to include and where this information can be located]. The module uses the outputs provided from MindFusionNet, to include: stress category, intensity score, and attention weights for different modalities, and retrieves matching items from the psychological knowledge base that include explanations, mental health guidelines, coping strategies, and next step recommendations. The use of established psychological research to ground the interpretation of raw outputs from the system enhances: (1) transparency, (2) reliability, (3) client understanding. Ulti- mately, the Medical Knowledge Base Integration Module acts as a bridge, between the artificial intelligence-based stress detection provided by the system, and a user's ability to apply that knowledge to their mental health.

The Prediction and Diagnosis Module uses the fused mul- timodal embeddings generated by MindFusionNet to deliver real-time stress assessment along with comprehensive inter- pretability. After receiving synchronized, noise-free inputs from NeuroClean and enriched hierarchical representations from DeepSenseX, the module outputs both a continuous stress-intensity score and a categorical stress level—low, medium, or high. Reliability is enhanced through uncertainty calibration, indicating how confident the system is in each prediction. It also generates modality-specific attention expla- nations that highlight key contributors to the detected stress state, such as physiological irregularities, subtle facial micro- expressions, or shifts in vocal tone. By linking these predic- tions to the integrated medical knowledge base, the module provides contextual behavioral insights, coping strategies, and personalized mental-wellness recommendations.

Fig. 2. Flow Diagram



This Module evaluates the Performance of the TriS-Mind Framework using a set of multimodal metrics designed specifically to evaluate the effectiveness of stress detection. Instead of evaluating how accurate the model was using just the "accuracy" measure, the model uses different classification metrics (precision, recall, F1-score, and confusion matrices).
The Reliability Scores generated from MindFusionNet's uncertainty calibration are reported to verify the clinical validity of the model. Cross modality Ablation Studies also provide a quantitative measure of the relative contribution of each modality used.
The temporal evaluation of the system is performed through the TEM component, which measures the degree to which the system captures the evolution of emotional patterns over time.

RESULT AND DISCUSSION

The analytical study of the TriS-Mind framework confirms the use of three phases within its multi-modal intelligence system to establish a sophisticated way to diagnose and understand the stress that individuals are experiencing through analysis of their text, video, audio, and biological signals. Clean data is processed through adaptive pre-processing by NeuroClean through elimination of specific signal modality noise, emotional timestamp alignment, maintenance of absolute variability in the expressive timing and the imputation of missing data through a multi-modal graph; this allows the data streams to be kept clean and synchronized for subsequent learning.
DeepSenseX extracts expressive, rich, emotion-aware representations of the various modalities via their respective encoders (i.e., using: a CNN-Transformer for Audio, a 3D Vision Transformer for Video, Frequency aware Temporal ConvNet for Physiological Data, and Emotional-BERT for Text), and through the integration of the multiple modalities through a cross modal projection layer and the refinement of the integrated representations by way of employing a Reliability-Based Harmonization Gate.

Accuracy

The TriS-Mind system's identification of mental stress level is provided through its accuracy measurement method, which considers all inputs received from multiple senses. This method evaluates the level of accuracy by determining how accurately (low, medium, high) the TriS-Mind System can predict an individual's stress state based on its input data. Accuracy is therefore an important indicator of the effectiveness of TriS-Mind across a wide range of environments.
The accuracy of TriS-Mind is calculated using the following formula:
(TP + TN) / (TP + FP + TN + FN)

Precision

Precision indicates how correctly the TriS-Mind system categorizes a particular level of stress versus how many times (out of all of those occurrences) it is guessing and therefore is critical to help reduce the number of false positives that could arise in detecting a person's mental state through multiple modes.
High levels of precision indicate the reliability of detection and identification through TriS-Mind as a result of NeuroClean filtering out erroneous background noises and DeepSenseX's application of reliability weighting to the feature extraction process.
Another way to increase precision for TriS-Mind is to use a mechanism such as cross-attention that focuses only on those modalities that are truly relevant to identifying someone in a stressed-out state. Through the combination of these three types of processes, the TriS-Mind system will only be triggered into alert mode (accumulate positive feedback) if strong, consistent patterns of evidence exist across modalities to indicate a stressed state. The calculation for precision will be TP (total predicted correct stress classifications) / (TP + FP (total predicted inaccurate classifications)).

Recall (Sensitivity/True Positive Rate)

In psychology, the measurement of recall is the most important indicator of how well the TriS-Mind system can identify when stress is accurately detected using various types of multimodal input data (including emotional/physiological signals). A high recall score reflects that a majority of an individual's genuine stressors have been captured and that the system is unlikely to miss any subtle emotional/physiological signals indicating the presence of stress.
When Multimodal Data Synchronized with the NeuroClean Platform is synchronized and missing/corrupted data is reconstructed, a significant increase in the overall recall measurement occurs due to the inclusion of previously excluded stress-related signals; thus, the reestablished connection helps to preserve genuine stress episodes. The DeepSenseX system (currently in beta testing) supports/reduces the number of false negatives which are commonly associated with stress detection by extracting rich associations at the emotional, acoustic, visual, and physiological levels and helping to highlight the unique stress patterns.
To aid in capturing all variations of the stress pattern over time, MindFusionNet has utilised both temporal memory and hierarchical Cross Attention modules to detect both short- and long-term changes in stress levels. By significantly minimising false negatives, the TriS-Mind system assists in ensuring that all genuine episodes of stress are detected, which is critical to supporting the clinical wellness community in ensuring that when a stress event occurs, timely support/interventions can occur.

F1-Score

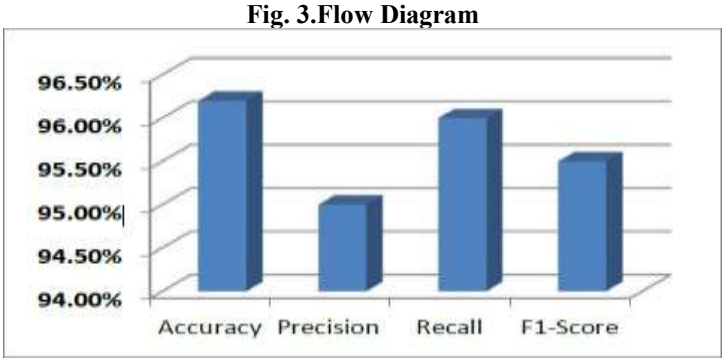
The F1 Score provides an indication of how effectively the TriS-Mind system has detected stress across different modalities by combining the Precision and Recall scores of a single test run and creating a harmonic average. Because mental states can involve subtle emotional cues through different modalities, the ability to accurately

identify stress (high Recall) and to avoid identifying faux or false positive stresses (high Precision) is especially vital for Mental States analysis purposes.
In addition, when considering the need for a multimodal framework such as TriS-Mind, which collects, processes, and analyzes data from text, audio, video, and physiological signals; the F1 Score is an important component in evaluating how well the model deals with multi-source data inconsistencies. TriS-Mind addresses both false-negatives and false-positives by leveraging the noise-adaptive signal preprocessing capabilities of NeuroClean, cross-modality embedding learning techniques employed in DeepSenseX, Hierarchical Fusion and temporal reasoning techniques used in MindFusionNet.
The successful combination and utilization of these technologies give rise to the high F1 Score associated with TriS-Mind's ability to predict and provide clinically-relevant and dependable predictions of stress in a variety of real-world scenarios.

Table 1. Comparison Table

Algorithm	Accuracy	Precision	Recall	F1-Score
copy	96.2%	0.95	0.96	0.955

Fig. 3.Flow Diagram



CONCLUSION

The TriS-Mind framework described here provides a sophisticated multimodal deep learning solution to detecting both the stress and emotional states of an individual by virtue of the integration with NeuroClean for effective pre-processing, DeepSenseX for hierarchical extraction of information across modalities as well as MindFusionNet for the final fusion and sense making of sensor data. The adaptive nature of the TriS-Mind framework allows it to be able to accurately process all four modalities (i.e., text, audio, video, and physiologic) while maintaining the inherent emotional context of the data collected so that clean, synchronized and contextually aligned representations can be produced for downstream analyses.
Additionally, providing interpretable attention maps, temporal emotion memory, uncertainty calibration and cross-modal projection further reinforce the TriS-Mind model's clinical utility and accuracy in prediction and classification of stress level(s), continuous intensity of stress, and reliability of classifications.
Finally, this framework will support tele-wellness platforms, decision support systems and real-time assessment of mental health conditions through its ability to enable continuous monitoring/information collection on stress level and an individual's overall mental well-being.

FUTURE WORK

Further development of the TriS-Mind framework will include the use of multiple forms of sensing to help assess emotion and cognitive (thinking) states. More signalling of a user's body will help researchers better understand the way a body reacts to stress or anxiety, i.e., the user will have more accurate information for each category.
The potential enhancement of the NeuroClean application through adaptive anomaly detection would assist in mitigating issues caused by unexpected noise; developing a self-supervised approach for DeepSenseX would provide the framework with a means to develop a more robust means for generating cross-modal representations (Cross-Modal Representation) using only limited amounts of labelled training data.
Researchers will also be able to investigate multiple, multi-task dimensions for MindFusionNet in an effort to create a single prediction for a number of aligned psychological constructs (for example, cognitive load, emotional exhaustion or anxiety) simultaneously.
The ultimate goal of deploying TriS-Mind on mobile and wearable devices will be possible using various types of privacy-preserving methods, such as federated learning, allowing for continuous monitoring and real-time tracking of mental well-being.

REFRENCES

1. Chen, H., et al. (2024). Deep-Emotion: A multimodal emotion recognition framework integrating speech, facial expressions, and EEG signals using attention-based fusion. IEEE Transactions on Affective Computing.
2. Tripathi, N., et al. (2023). A multimodal stress detection model using EMD-based EEG feature extraction, MFCC speech analysis, and BiLSTM fusion for temporal correlation learning.
3. Zhang, S., et al. (2022). A multimodal emotion recognition system using LSTM-based physiological signal analysis and 3D-CNN spatiotemporal facial feature extraction with late fusion.

4. Kumar, A., et al. (2023). A multimodal stress recognition framework integrating EEG–ECG signals and facial emotion features using WPD, HOG, and Deep Belief Network fusion.
5. Li, T., et al. (2024). A transformer-based multimodal fusion network using GCN-encoded EEG, ResNet-50 audio features, and BiLSTM video analysis for robust emotion and stress recognition.
6. Zhang, J., Yin, Z., Chen, P., Nichele, S. (2020). Emotion recognition using multimodal data: A systematic review. IEEE Access, 8, 133813–133831.
7. Tripathi, S., Acharya, S., Sharma, R. D., Mittal, S., Bhattacharya, S. (2017). Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. AAAI Workshops.
8. Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. ACL.
9. Kim, J., André, E. (2008). Emotion recognition based on physiological changes in music listening. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(12), 2067–2083.
10. Akçay, M. B., Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, and classifiers. Speech Communication, 116, 56–76.