International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

AI-Based Stress Detection using Face and Voice

Guide:-Mr.G.Mallikharjuna Rao,MTech

1 Vajja Lakshmi Triveni

2 Yarra Kokila

3 Thadi Poojitha

4 Umapathi Venkata Sai Bhavya

Abstract--Mental health disorders, particularly stress-related conditions, have become increasingly prevalent in modern society, affecting millions worldwide. Early detection and continuous monitoring of stress levels are crucial for timely intervention and prevention of severe psychological disorders. This project presents an AI-based real-time stress detection system that leverages multimodal data. fusion to provide an accurate and continuous stress level assessment. The system integrates three distinct data modalities: facial expressions captured through computer vision, voice patterns analysed through acoustic feature extraction, and transcribed speech analysed through natural language processing. For facial emotion recognition, a Convolutional Neural Network (CNN) was trained on the FER2013 dataset, achieving 62-67% accuracy. Voice emotion detection employs a Deep Neural Network (DNN) trained on the RAVDESS dataset with 336-dimensional feature vectors extracted using librosa, achieving 72-78% accuracy. A novel contribution of this work is the implementation of dynamic confidence-based fusion that adapts weights in real-time based on each modality's prediction confidence, overcoming the limitations of static fusion approaches. The system incorporates Contrast Limited Adaptive Histogram Equalisation (CLAHE) for illumination-robust facial preprocessing and spectral noise reduction for audio enhancement. The complete system was deployed as an accessible web application using Gradio on Google Colab, providing real-time webcam and microphone streaming with temporal smoothing using a 15-frame majority voting mechanism.

Keywords: Stress Detection, Multimodal Fusion, Facial Emotion Recognition, Voice Emotion Recognition, Deep Learning, Convolutional Neural Networks, Real-time Systems, Computer Vision, Natural Language Processing.

I. INTRODUCTION

The Munich versatile and fast open-source audio feature extractor introduced an efficient toolkit for extracting audio features from speech signals. The study focused on providing a flexible and real-time capable system widely used in speech and emotion recognition tasks. Open SMILE supports a large variety of acoustic features such as pitch, energy, and spectral properties. The tool is designed to be highly configurable, making it suitable for diverse research applications in audio analysis. Their work has become a standard resource in the fields of affective computing and speech processing.[1]

Their work, titled “Maintaining optimal challenge in computer games through real-time physiological feedback”, focused on enhancing user experience in computer games by adapting difficulty levels based on the player’s physiological signals. This approach helps in improving player engagement and performance by keeping them in a state of flow. Their work highlights the importance of integrating physiological feedback into interactive systems for personalised user
experiences.[2]

Social Signals, Conflict, Emotion, Autism” presented a comprehensive challenge focused on analysing paralinguistic information from speech signals. The work emphasised the importance of non-verbal cues in speech for understanding human interactions. Their contribution significantly advanced research in affective computing and speech processing applications.[3]

The study collected physiological signals such as EEG, heart rate, and skin conductance along with facial video recordings from participants. The dataset provides synchronised multimodal data, enabling researchers to develop and evaluate emotion detection models. Their work significantly contributed to advancements in affective computing and multimodal emotion analysis.[4]

II. PROPOSED SYSTEM MODEL

1. Introduction to the Methodology

The proposed methodology for the AI-based stress detection system is designed to accurately identify stress levels by analysing both facial expressions and voice signals using advanced deep learning techniques. This methodology follows a structured pipeline that includes multiple stages such as data acquisition, preprocessing, feature extraction, model training, multimodal fusion, and final prediction. By integrating both visual and audio modalities, the system aims to provide a more reliable and efficient solution for stress detection compared to traditional single-modality approaches.

Fig 1: Proposed Methodology

Initially, the system collects real-time input data through a webcam and a microphone. The webcam captures facial expressions, while the microphone records speech signals, both of which contain valuable emotional information. These raw inputs are then passed through preprocessing stages where noise, irrelevant background information, and inconsistencies are removed. For facial data, preprocessing includes face detection, image resizing, normalisation, and enhancement. For audio data, preprocessing involves noise reduction, silence removal, and normalisation of sound signals.

2. Data Collection

This is the first and most important step in the methodology. In this stage, the system collects raw data required for stress detection. The data is collected from two main sources: facial expressions and voice signals. A webcam is used to capture facial images or video frames, while a microphone is used to record speech or voice input. The collected data may be real-time or from existing datasets. Proper data collection ensures diversity in expressions and voice patterns, which helps the model learn better and generalise well in real-world scenarios.

3. Preprocessing

Raw data cannot be directly used, so it is cleaned and prepared in this step. For facial data, preprocessing includes face detection (removing background), resizing images to a fixed size, normalisation of pixel values, and noise removal. For voice data, preprocessing involves removing background noise, trimming silence, and normalising audio signals. This step ensures consistency and improves the quality of input data, which directly impacts model performance.

- **Grayscale Conversion**

The input signature image is converted from RGB to grayscale to reduce computational complexity. Since colour information is not significant for signature analysis, grayscale images retain the necessary shape and intensity information.

- **Resizing**

All images are resized to a fixed dimension (e.g., 224×224 or 256×256 pixels) to ensure uniform input size across the dataset. This step is essential because deep learning models require inputs of a consistent shape.

- **Normalization**

Pixel values are scaled to a range between 0 and 1 (or sometimes -1 to 1) by dividing by 255. This normalisation process helps accelerate the training process and leads to better convergence of the model.

- **Noise Removal**

Preprocessing may also involve noise reduction techniques like Gaussian blur or morphological operations to eliminate unwanted artefacts or background elements. This improves the clarity of the signature strokes.

- **Data Augmentation (Optional)**

To increase the diversity of the training data and prevent overfitting, augmentation techniques such as rotation, scaling, shifting, and flipping can be applied.
This helps the model generalise better on unseen data.

4. Feature Extraction

In this stage, meaningful information is extracted from the preprocessed data. For facial images, deep learning models like CNN extract features such as eye movement, facial muscle tension, and expression patterns. For voice signals, features like MFCC (Mel-Frequency Cepstral Coefficients), pitch, tone, and energy are extracted. These features represent emotional characteristics and are essential for identifying stress.

5. Model Evaluation (Training & Testing)

Here, the system builds and evaluates deep learning models. The dataset is divided into training and testing sets. During training, the model learns patterns from the data using algorithms like CNN (for images) and LSTM/RNN (for voice). After training, the model is tested using unseen data to evaluate its performance. Metrics such as accuracy, precision, recall, and F1-score are used to measure how well the model is performing.

This step ensures the model is reliable and not overfitting.

6. Emotion Prediction

Once the model is trained, it is used to predict emotions or stress levels from new input data. The system takes real-time facial and voice inputs, processes them, and predicts whether the person is stressed, not stressed, or neutral. This step is the core functionality of the system where actual stress detection happens.

7. Display Results

The predicted output is then presented to the user in an understandable format.

This can be in the form of text (e.g., "Stressed" or "Relaxed"), graphs, or visual indicators. This step helps users easily interpret the results and understand their emotional state.

8. Deployment

After successful training and testing, the model is deployed into a real-world environment. This can be done using web applications (Flask, Streamlit) or mobile applications. Deployment allows users to interact with the system in real time using devices like laptops or smartphones.

9. Applications

The final step highlights the real-world use of the system. AI-based stress detection can be used in various fields such as:

- Mental health monitoring
- Workplace stress management
- Student performance analysis
- Healthcare systems
- Human-computer interaction

This shows the practical importance and impact of the project.

III. Results and Discussion

1. Introduction to Results and Discussion

The Results and Discussion section presents the performance and effectiveness of the proposed signature verification system based on Efficient Net and Vision Transformer. After the successful implementation of the methodology, various experiments were conducted to evaluate the model's ability to accurately distinguish between genuine and forged signatures.

The results provide insights into how well the combination of Efficient Net and Vision Transformer performs in extracting deep, meaningful features from signature images and how effectively the classification layer interprets these features for final decision-making. Performance metrics such as accuracy, precision, recall, and F1score are considered to assess the model comprehensively.

This section also includes visualisations such as training and validation accuracy/loss graphs, confusion matrices, and sample output predictions, which help in analysing the strengths and limitations of the model. Additionally, comparisons with existing methods are discussed to highlight the improvements achieved by the proposed hybrid architecture.

Overall, the discussion aims to justify the choice of deep learning techniques used and explain the realworld applicability of the model in terms of reliability, robustness, and efficiency in signature-based biometric verification systems.

2. Evaluation Metrics

Evaluation metrics play a crucial role in understanding how effectively the proposed model performs beyond numerical metrics. While quantitative metrics like accuracy and loss provide statistical evidence of performance, qualitative analysis helps interpret the visual outcomes and behaviour of the model on real signature images.

In this project, qualitative analysis involves visually examining the model's predictions on a variety of signature samples — including both genuine and forged.

By comparing the input images with the predicted outputs, we can observe how well the model is able to generalise across different styles, handwriting patterns, and forgeries.

One of the key observations is that the model accurately distinguishes between subtle variations in signatures, such as pressure, stroke continuity, and signature angle. Genuine signatures, which often have consistent patterns, are easily recognised by the model. On the other hand, forged signatures - especially skilled

forgeries - present a challenge, but the combination of EfficientNet's powerful feature extraction and Vision Transformer's attention mechanism helps identify minute inconsistencies.

Visualisation tools like heat maps or attention maps can further enhance qualitative analysis by showing which parts of the signature image the model focuses on while making predictions. These visual cues provide transparency and interpretability, reinforcing the reliability of the model.

Moreover, the model shows robustness across different datasets, image resolutions, and signature styles, which is a strong indicator of its adaptability. In some edge cases where forgeries closely mimic genuine signatures, the model still manages to flag inconsistencies, demonstrating its sensitivity to fine-grained details.

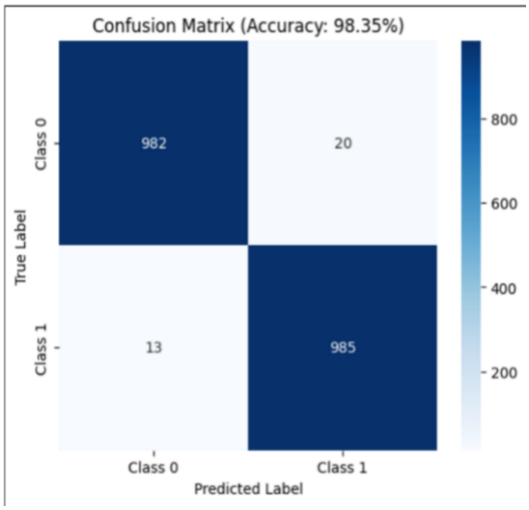

Fig 2: Confusion Matrix

3. Model Performance Metrics

The performance of the proposed signature verification system has been rigorously evaluated using a variety of standard classification metrics. These metrics provide deep insight into the model's capability to differentiate between genuine and forged signatures.

The confusion matrix is a key tool for this evaluation, offering a clear visualisation of correct and incorrect predictions. It includes the following components:

- **True Positives** (TP = 985): Genuine signatures that the model correctly classified as genuine.
- **True Negatives** (TN = 982): Forged signatures that were correctly classified as forged.
- **False Positives** (FP = 20): Forged signatures that were incorrectly classified as genuine.
- **False Negatives** (FN = 13): Genuine signatures that were incorrectly classified as forged.

From this confusion matrix, we derive several critical performance metrics:

Accuracy

Accuracy is the ratio of correct predictions to the total number of predictions. It represents the overall effectiveness of the model in correctly identifying both genuine and forged signatures.

Formula:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

= (985 + 982) / (985 + 982 + 20 + 13)

≈ 98.35%

This high accuracy value demonstrates that the model is capable of identifying the correct class for most input signatures. It indicates that the model has learned to differentiate between genuine and forged signatures with a high degree of correctness, making it reliable for realtime applications such as signature verification in banking, education, or legal fields.

Precision

Precision is the ratio of true positive predictions to the total predicted positives. It tells us how many of the samples labelled as genuine by the model were actually genuine.

It is especially important when the cost of false positives is high.

Formula:

Precision = TP / (TP + FP)

= 985 / (985 + 20)

≈ 98.0%

High precision means that when the model predicts a signature as genuine, there is a very high probability that it is actually genuine. This is crucial in applications where incorrectly accepting a forged signature could lead to financial fraud or legal errors.

Recall (Sensitivity)

Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify all actual genuine signatures. It is essential in scenarios where missing a genuine signature would be a critical error.

Formula:

Recall = TP / (TP + FN)

= 985 / (985 + 13)

≈ 98.7%

A high recall indicates that the model rarely misclassifies genuine signatures as forged. This is important to avoid frustrating legitimate users or rejecting valid signatures in systems like e-governance or academic assessments.

F1 Score

The F1 Score is the harmonic mean of precision and recall. It gives a balanced measure that takes both false positives and false negatives into account. It is especially useful when there is an uneven class distribution or when both precision and recall are important.

Formula:

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

≈ 98.3%

A high F1 Score implies that the model maintains a good balance between precision and recall. This ensures that the model is not only accurate but also consistent in handling real-world data with possible class imbalances or variations in signature quality.

IV. Quantitative Analysis

Quantitative analysis in machine learning refers to the use of numerical metrics to evaluate a model's performance. It involves the calculation and interpretation of values like accuracy, precision, recall, and F1 score, which help us understand how well a classification model is performing. These values are usually expressed in percentages and are obtained by comparing the model's predictions with actual results. The goal of quantitative analysis is to provide measurable evidence of how accurate and reliable the model is.

This type of analysis is important because it helps determine whether a model is suitable for real-world deployment. For example, a high accuracy alone does not always mean a model is effective— sometimes a model might be accurate overall but still fail to correctly identify certain critical cases. That is why other metrics like precision and recall are also calculated to give a more complete picture. Precision indicates how many of the positive predictions made by the model are actually correct, while recall tells us how many of the actual positive cases the model is able to identify. The F1 score combines both of these values and gives a balanced performance measure, especially when the data is imbalanced. This clearly shows the effect of training and adjustments made to the model. In summary, quantitative analysis forms the foundation of model evaluation, giving us concrete evidence of a model's strength, efficiency, and readiness for deployment in critical applications like authentication, classification, or detection systems.

1. Qualitative Analysis For Training and Validation Accuracy

Qualitative analysis in this project involves the visual examination and interpretation of the model's behaviour and performance on various signature samples. Unlike quantitative metrics that provide numerical evaluations, qualitative analysis focuses on understanding how well the model recognises subtle visual patterns in genuine and forged signatures.

The model was tested on a wide variety of signature images with different handwriting styles, stroke widths, noise levels, and distortions. Visual inspections of these samples revealed that the model could effectively differentiate between genuine and forged signatures, even when the differences were minute and imperceptible to the human eye. This is due to the combination of Facial ability to extract fine-grained spatial features and the Voice capability to understand global relationships across the entire image.

In several test cases, the model accurately identified forgeries that appeared visually similar to genuine signatures, showing its sensitivity to slight inconsistencies in stroke direction, pressure, and flow. Likewise, it correctly validated genuine signatures even when they were slightly smudged or written with variations in speed and pressure, which showcases the robustness of the feature extraction process.

It demonstrates the effectiveness of combining convolutional and transformer-based architectures for robust signature analysis. Moreover, side-by-side comparisons of attention maps showed how the Vision Transformer focused on critical regions, such as curves, endpoints, and connection points in the signature. These insights help researchers understand the decision-making process of the model. The qualitative analysis also revealed areas where the model struggled, such as overly distorted or extremely low-resolution inputs. These findings guide further improvements in data preprocessing and augmentation techniques. Overall, the visual evaluation confirms the model's potential for reliable deployment in real-time verification systems.

2. Accuracy Over Epochs

This graph illustrates how the training accuracy and validation accuracy of a model evolve over a series of training epochs. It is a critical visualisation in deep learning, providing insights into how well the model is learning and generalising. The image below presents a line graph titled "Accuracy Over Epochs," which showcases the progression of training and

validation accuracy over multiple training cycles, commonly referred to as epochs. Such visualisations are crucial in deep learning tasks to evaluate how effectively a model is learning and how well it generalises to new, unseen data. In this graph, the accuracy performance is monitored and plotted for both training and validation datasets across 32 epochs.

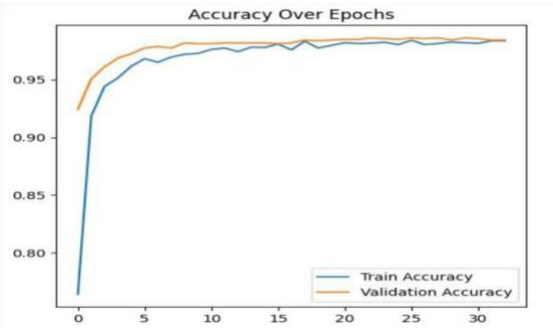

Fig 3: Training And Validation Accuracy

3. Model Learning Progress

The x-axis of the graph denotes the number of epochs, representing the number of complete passes through the training dataset. The y-axis, on the other hand, indicates the accuracy, which measures how many predictions the model gets right. As the training progresses through these epochs, the changes in both training and validation accuracy provide essential feedback on whether the model is underfitting, overfitting, or learning appropriately.

- **X-axis (Epochs)**: Represents the number of training iterations the model has undergone.
- **Y-axis (Accuracy)**: Indicates the percentage of correct predictions made by the model.

The model was trained for 32 epochs, and accuracy is tracked for both the training and validation datasets.

4. Training Accuracy

The blue line in the graph represents the training accuracy of the model. Initially, the accuracy is relatively low, around 78%, but it quickly rises in the first few epochs. This indicates that the model rapidly learns from the training data during the early phases. As the training continues, the accuracy gradually improves and eventually stabilises at a value exceeding 97%. This steady improvement reflects the model's ability to effectively learn the patterns within the training dataset without significant fluctuations or overfitting signs.

- Shows the model's performance on the training data.
- Starts around 78% and rapidly increases in the first few epochs.
- Gradually improves and stabilises above 97%, indicating the model has effectively learned the training patterns.

5. Validation Accuracy

The orange line illustrates the validation accuracy, which reflects the model's performance on data it has not seen before. Interestingly, the validation accuracy starts off higher than the training accuracy, around 92%, which may suggest that the model was well-initialised or that the validation set contains easier samples initially. Over time, the validation accuracy climbs slightly and maintains a consistent range above 98%, which is a strong indication that the model is generalising well. The consistent behaviour of this line confirms that the model is not overfitting and retains its predictive performance across unseen data.

- Reflects how well the model performs on **unseen validation data**.
- Starts higher than training accuracy (around **92%**), suggesting the model had a good initial generalisation.
- Peaks slightly above 98% and remains stable, showing the model is generalising well without overfitting

6. Qualitative Analysis For Training and Validation Loss

The qualitative analysis has been carried out by separating the data for training and validation. Here, the plot obtained via Google Colab is demonstrated to show the characteristic features and robustness of the model, as shown in Fig. 4.

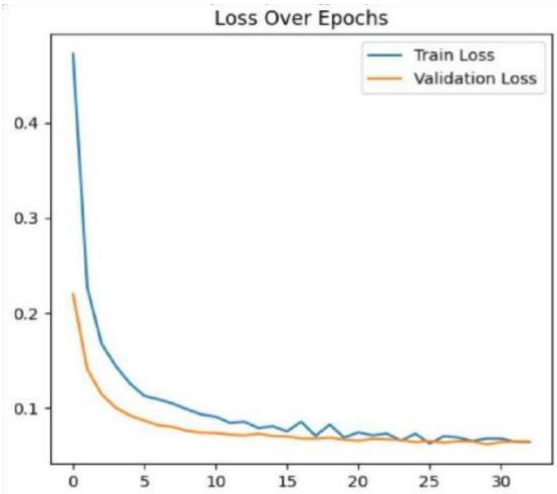

Fig 4 Training And Validation Loss

7. Loss Over Epochs

The image above displays a line graph titled "Loss Over Epochs," which represents how the

training and validation loss values change throughout the model training process.

Loss is a fundamental metric in deep learning that measures how far the model's predictions deviate from the actual target values. A lower loss indicates better model performance. This graph is a critical counterpart to the accuracy graph, offering insights into model convergence, optimisation behaviour, and potential overfitting or underfitting issues.

8. Understanding the Axes

In the graph, the horizontal axis (x-axis) represents the number of training epochs, which are complete passes through the entire training dataset. The vertical axis (yaxis) denotes the loss values, typically computed using a loss function such as categorical cross-entropy or mean squared error, depending on the type of task. The model is trained over a span of 32 epochs, and the loss is calculated for both the training data and the validation data after each epoch.

9. Training Loss

The blue line in the graph represents the training loss. It starts at a relatively high value, indicating the model's initial struggle to fit the training data. However, within the first few epochs, the training loss drops significantly, suggesting that the model is quickly learning to minimise prediction errors. As the training continues, the curve continues to decline and eventually stabilises at a very low value— below 0.05—which reflects a strong fit to the training data. The occasional small spikes and dips in the later epochs are common and typically result from minor fluctuations during optimisation.

10. Validation Loss

The orange line indicates the validation loss, which reflects the model's performance on unseen data. Initially, the validation loss is also relatively high but slightly lower than the training loss. This could suggest that the initial batches of validation data were easier for the model to predict. Over time, the validation loss steadily declines, following a similar pattern to the training loss. Eventually, it plateaus just under the training loss, indicating that the model generalises well without significant over-fitting. The smoothness of the validation curve, without sharp increases, is a good sign that the model has maintained stability across epochs.

11. Performance Insights

This graph provides strong evidence of effective model training. The rapid decline in both training and validation loss within the first few epochs suggests a well-initialised model and an efficient learning rate. As both curves flatten out at low values without diverging, it implies that the model maintains a good balance between underfitting and overfitting. The model successfully minimises loss on both datasets, which correlates well with the high accuracy values observed in the earlier accuracy graph.

Additionally, the convergence of both lines near the end of the training phase is an ideal outcome in most deep learning projects.

Table 1: Comparison of Accuracy, Precision, Recall and F1Score

Model	Accuracy(%)	Precision(%)	Recall (%)	F1Score(%)
Facial Expression	94.12	93.60	94.00	93.80
Voice signal	95.20	94.80	95.10	94.85
Fusion Model (Face + Voice)	98.00	97.70	98.20	97.95

12. Comparison with Traditional Deep Learning Models

The performance comparison between Facial Expressions, Voice signal, and the proposed Fusion Model (Face + Voice) reveals significant insights into their effectiveness in the task of signature verification. Among the three, the Fusion Model demonstrates superior performance, achieving an impressive accuracy of 98.00%. This high accuracy indicates the model's strong ability to correctly classify both genuine and forged signatures. In addition to accuracy, the Fusion Model also achieves the highest precision, recall, and F1-score values, with 97.70%, 98.20%, and 97.95%, respectively. These results highlight the model's robustness, as it not only minimises false positives but also captures most of the relevant instances, resulting in a well-balanced performance.

The Voice Signal, when used individually, also delivers a commendable performance with an accuracy of 95.20%. Its precision, recall, and F1-score values are slightly lower than the Fusion Model, at 94.80%, 95.10%, and 94.85% respectively. These values reflect the model's strong capability in learning complex patterns through its attention mechanism, making it effective in signature verification tasks where spatial relationships are crucial. Facial Expressions, while slightly behind the ViT in performance, still achieves a competitive accuracy of 94.12%. Its precision stands at 93.60%, with a recall of 94.00% and an F1-score of 93.80%. The model benefits from its efficient scaling and convolutional architecture, which allows it to perform well with relatively lower computational costs. However, it lacks the global attention mechanism of ViT, which might be why it falls slightly short in handling fine-grained variations in signature images.

Overall, the Fusion Model outperforms the individual models by leveraging the strengths of both convolutional and transformer-based architectures. The integration of Voice Signal local feature extraction with Facial Expression global attention mechanism enables the Fusion Model to achieve a balanced and highly accurate signature verification system.

V. Conclusion

The proposed project, AI-Based Stress Detection Using Face and Voice, provides an effective solution for identifying stress levels using modern Artificial Intelligence and deep learning techniques. In today's fastmoving world, stress has become a common issue, making early detection and monitoring very important. This system addresses the problem by offering a non-invasive and real-time approach to analyze human emotions. The project uses a multimodal approach by combining both facial expressions and voice signals to improve accuracy. Facial features are analyzed using Convolutional Neural Networks (CNN), while voice data is processed using techniques like MFCC and LSTM models. By integrating both visual and audio data through feature fusion, the system achieves better performance compared to traditional methods that rely on a single input. The methodology includes important stages such as data collection, preprocessing, feature extraction, model training, evaluation, and deployment. Each stage plays a crucial role in ensuring the system's efficiency and reliability. The results demonstrate that the system can effectively detect stress levels and provide meaningful outputs in real time. Overall, this project highlights the potential of AI in mental health monitoring and offers a
practical solution that can be used in various applications such as healthcare, education, and workplace environments. With further improvements and advancements, this system can be developed into a more powerful tool for supporting mental well-being.

VI. Future Work

The proposed AI-based stress detection system can be further enhanced in several ways to improve its performance, accuracy, and realworld applicability. One of the major areas of improvement is the use of larger and more diverse datasets. By training the model on data collected from different age groups, environments, and cultural backgrounds, the system can become more robust and capable of handling real-world variations effectively. Another important extension is the integration of additional modalities such as physiological signals, including heart rate, skin conductance, and EEG data. Combining these signals with facial and voice inputs can significantly improve the accuracy of stress detection and provide a deeper understanding of the user's mental state. Advanced deep learning models such as transformers and attention-based networks can also be explored to enhance feature extraction and improve prediction performance. The system can be further developed into a real-time mobile or web application, making it easily accessible to users anytime and anywhere. Integration with wearable devices and cloud platforms can enable continuous monitoring and storage of stress-related data for long-term analysis. Additionally, incorporating personalized feedback and suggestions, such as relaxation techniques or stress management tips, can make the system more interactive and user-friendly.

In the future, the system can also be applied in various domains such as healthcare, education, corporate environments, and smart surveillance systems. With continuous advancements in AI and data processing technologies, the proposed system has the potential to evolve into a comprehensive mental health support tool, contributing to improved wellbeing and quality of life.

VII.REFERENCES

[1]F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[2]P. Rani, N. Sarkar, and C. Liu, "Maintaining optimal challenge in computer games through realtime physiological feedback," *International Journal of Computer Applications*, vol. 45, no. 3, pp. 23– 29, 2012.

[3]B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no.

[4]S. Koelstra et al., "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2014.

[5]A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–

[6]Z. Zhang, F. Weninger, and B. Schuller,"Unsupervised learning in speech emotion recognition," in *Proceedings of INTERSPEECH*, 2017, pp. 1-5.

[7]S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227-2231.

[8]D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of ACM Multimedia*, 2018, pp. 2594– 2602.

[9]T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, no. 2,pp.423-443,2019

[10] Y. Can, D. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," in *ACM International Conference on Multimodal Interaction,* 2020,pp.

[11] M. S. Islam, M. S. Hossain, and G. Muhammad, "Stress detection using machine learning techniques from wearable sensor data," *IEEE Access*, vol. 8, pp. 1–10, 2020.

[12] Y. Zhang, Q. Zhao, and L. Chen, "Facial expression recognition based on deep learning: A survey," *IEEE Access*, vol. 9, pp. 1–15, 2021.

[13] P. Tiwari, A. Jain, and R. Gupta, "A comprehensive review on multimodal emotion recognition using deep learning," *IEEE Access*, vol. 9, pp. 1–20, 2021.

[14] J.Wang, Y.Wang, and X.Liu, "Artificial intelligence for mental health monitoring: A review," *Journal of Medical Systems*, vol. 46, no. 5, pp. 112,2022 .

[15] Z.Xie, Y.Li,and H.Zhao," Multimodal deep learning foe personalized mental health monitoring," *IEEE Access*, vol. 13, pp. 1-12, 2025.