International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert/2024/v11/i3/v11i306
Volume 11, Issue 3, March 2024
© 2024, IJCERT All Rights Reserved

Research Paper

Speech-Based Emotion Recognition and PTSD Detection through Machine and Deep Learning

[1*]K Islam, [2]Z ElSayed
[1*]*Professor in Information Analysis and Information Retrieval, Mansoura University, Egypt*
[2] *Faculty of Medicine. Zagazig University, Zagazig, Egypt*

**Corresponding Author(s): islam.k786@mans.edu.eg*

Received: 11/11/2023, Revised: 19/02/2023, Accepted:23/03/2024 Published:30/03/2024

Abstract: This study investigates the potential of machine and deep learning algorithms for Speech Emotion Recognition (SER) and Post-Traumatic Stress Disorder (PTSD) detection through speech analysis. Traditional diagnostic methods for PTSD, which are often subjective and time-consuming, are in contrast with the automated capabilities offered by these algorithms, enabling early detection through the identification of specific speech patterns. Utilizing the RAVDESS Emotional Speech Audio dataset alongside PTSD-specific recordings, this study applies preprocessing techniques such as noise reduction and normalization to enhance the quality of the speech data. Feature extraction is performed by focusing on acoustic, linguistic, and temporal features that capture variations in the pitch, intonation, and speech rate. Both machine learning models, including Support Vector Machines (SVMs) and Random Forests, and deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have been developed and compared. Experimental results indicate that deep learning models achieve up to 91% accuracy in SER and 89% accuracy in PTSD detection, significantly outperforming traditional machine learning methods. These findings demonstrate the efficacy of multimodal integration in improving diagnostic capabilities, particularly through a combination of speech, text, and physiological data. However, the study acknowledges the limitations in generalizability across diverse populations and the practical challenges of deploying these models in real-world applications. Future work will focus on expanding the datasets to include a wider range of demographic and cultural variations, enhancing real-time monitoring capabilities, and refining model interpretability to ensure reliable performance in various contexts.

Keywords: speech-based emotion recognition, PTSD detection, machine learning, deep learning, feature extraction, multimodal integration.

1.Introduction

Human emotions play a crucial role in social interaction and well-being. Recognizing emotions accurately can be invaluable in various applications, including healthcare, human-computer interaction, and customer service. Speech offers a rich source of information for emotion recognition, as vocal characteristics such as pitch, intonation, and prosody often reflect emotional states.Speech Emotion Recognition (SER) is a dynamic field within affective computing focused on the identification of human emotions from speech using computational techniques. By analysing vocal characteristics such as pitch, intonation, rhythm, and speech rate, SER has evolved from traditional handcrafted feature extraction and statistical models to leveraging advanced machine and deep learning algorithms, which can automatically learn relevant features from extensive datasets. This evolution has significantly enhanced the accuracy of emotion classification by capturing intricate speech patterns and nuances [1], [2]. Accurate emotion recognition is crucial across several domains. In healthcare, SER provides an objective tool for diagnosing and managing mental health conditions such as depression and Post-Traumatic Stress Disorder (PTSD), offering real-time monitoring and supporting therapeutic interventions [3]. In human-computer interaction, SER enables the development of empathetic systems that respond to users' emotional states, enhancing user experience and engagement. Additionally, in customer service, SER improves automated systems by adapting responses based on emotional cues, leading to better customer satisfaction and service quality [4]. These applications illustrate the versatility and transformative potential of SER in healthcare, human-computer interaction, and customer service.

This research explores the possibility of utilizing speech analysis for PTSD detection as a potential complementary tool to traditional diagnostic methods. We investigate the ability of machine learning and deep learning algorithms to identify speech patterns associated

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

with PTSD, potentially aiding in early detection and diagnosis.

This paper presents a comprehensive investigation into speech-based emotion recognition and PTSD detection. We explore various machine learning and deep learning techniques for both tasks, analyzing their effectiveness in identifying emotions from speech and potentially detecting PTSD markers. The research aims to contribute to the advancement of both SER and PTSD diagnosis, demonstrating the potential of speech analysis as a valuable tool in the realm of mental health.

This study presents several key contributions to the field of speech-based emotion recognition and PTSD detection:

- **Speech-based Emotion Recognition and PTSD Detection:** Investigates the potential of using machine and deep learning algorithms for both automatic speech emotion recognition (SER) and post-traumatic stress disorder (PTSD) detection through speech analysis.
- **Machine and Deep Learning for SER:** Explores the effectiveness of machine learning and deep learning techniques in identifying emotions from speech, contributing to the advancement of SER research.
- **PTSD Detection through Speech Analysis:** Examines the possibility of utilizing speech patterns to aid in PTSD detection, potentially serving as a complementary tool to traditional diagnostic methods.
- **Enhanced Mental Health Diagnosis:** Aims to contribute to improved mental health diagnosis by demonstrating the potential of speech analysis for both SER and PTSD detection.

Following the introduction, the paper includes Section 2 on literature review, comparing methods in SER and PTSD detection. Section 3 covers the methodology, including data processing, feature extraction, and model development. Section 4 details experiments and results, comparing model performances. Section 5 discusses limitations such as generalizability and real-world applicability. Section 6 concludes with findings and proposes future work for enhancing model effectiveness.

2. Literature Review

2.1 Emotion Recognition in Speech

The development of Speech Emotion Recognition (SER) has undergone significant evolution, beginning with basic handcrafted feature extraction methods and advancing to sophisticated deep learning techniques. Initially, SER relied on classical machine learning approaches, where features such as pitch, energy, and spectral properties were manually extracted and used in models like Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) for emotion classification. These early methodologies provided foundational insights but often struggled with the complexity and variability inherent in human speech [5], [6].

As the field progressed, the advent of deep learning revolutionized SER by enabling the automatic learning of features directly from raw speech data. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including their advanced variants like Long Short-Term Memory (LSTM) networks, allowed for the capture of more intricate patterns in speech, leading to substantial improvements in emotion recognition accuracy. This transition marked a significant shift from rule-based and shallow learning methods to more robust, data-driven approaches [7], [8]. Presently, SER research continues to explore various neural network architectures and multimodal integration strategies, focusing on improving generalizability and performance across diverse datasets and real-world applications [9].

2.2 PTSD Detection

Traditional diagnostic methods for Post-Traumatic Stress Disorder (PTSD) primarily involve clinical interviews and self-reported questionnaires, which, while useful, are often limited by subjectivity and the time required for thorough evaluation. These methods typically rely on the identification of symptomatic patterns such as intrusive memories, avoidance behaviors, and hyperarousal following traumatic events [10]. However, recent research has begun to explore the potential of speech as a valuable tool in mental health assessment, particularly for PTSD.

Speech analysis for PTSD detection aims to identify vocal biomarkers associated with the disorder, including variations in tone, speech rate, and prosody that may reflect underlying emotional distress or cognitive alterations. Studies have utilized machine learning and deep learning techniques to analyze these speech patterns, demonstrating promising results in identifying PTSD-related features that might not be easily detectable through traditional methods [11]. This approach offers a non-invasive and objective means of complementing existing diagnostic procedures, potentially facilitating earlier and more accurate detection of PTSD [12], [13].

2.3 Intersection of SER and PTSD Detection

Combining Speech Emotion Recognition (SER) with PTSD detection presents unique opportunities and challenges. Both fields leverage speech analysis but focus on different aspects: SER primarily aims at classifying general emotional states, while PTSD detection seeks to identify specific pathological speech patterns associated with trauma. Techniques such as deep learning can be applied to both areas, allowing for the integration of emotional and diagnostic analysis from speech data [14].

One major challenge in combining SER and PTSD detection lies in distinguishing between general emotional expressions and PTSD-specific speech markers. Emotional states may vary widely across individuals and contexts, whereas PTSD-related speech patterns might be more subtle and complex. Additionally, developing models that can accurately classify both emotion and PTSD markers requires extensive, high-quality datasets that capture the diversity of human speech and the nuanced characteristics

of PTSD [15], [16]. Despite these challenges, the integration of SER with PTSD detection holds significant promise for enhancing the accuracy and efficiency of mental health assessments, potentially leading to better clinical outcomes [17].

Table 1: Comparative Analysis of Multimodal Emotion and PTSD Detection Studies

Reference	Objective & Methods	Key Findings
Muzammel et al. [5]	Depression recognition using end-to-end multimodal deep neural networks (speech, text, visual).	85% accuracy in depression recognition using combined modalities.
Othmani et al. [6]	PTSD diagnosis with machine learning-based video and EEG analysis.	Combined video and EEG data achieved >80% accuracy in diagnosing PTSD.
Shoumy [7]	Emotion recognition via multimodal data (text, audio, visual) with data augmentation and fusion.	90% accuracy using enhanced data fusion techniques.
Kuttala et al. [9]	Stress detection with hierarchical CNNs using multimodal data (physiological, behavioral).	Hierarchical CNNs reached 85% accuracy in stress detection.
TJ et al. [17]	Depression analysis with D-ResNet-PVKELM on multimodal data (speech, text, physiological).	Achieved ~88% accuracy in multimodal depression

This Table 1 summarizes key studies that explore the use of multimodal data and deep learning methods for emotion recognition and PTSD detection. The studies utilize various combinations of data modalities such as speech, text, visual, physiological, and EEG to enhance detection accuracy. Advanced machine learning techniques including deep neural networks and hierarchical CNNs are applied, achieving accuracy rates ranging from 85% to 90%, demonstrating the effectiveness of multimodal approaches in improving diagnostic capabilities.

2.4 Research Gaps

- Insufficient seamless fusion of diverse data modalities for unified emotion and PTSD detection models.
- Lack of validation across diverse linguistic, demographic, and cultural backgrounds.
- Limited research on continuous, real-time monitoring systems for dynamic environments.
- Under exploration of emotional context and individual variability's impact on detection accuracy.
- Inadequate attention to mitigating biases and ethical implications in data and algorithms.
- Need for more interpretable models to understand decision-making in deep learning systems.
- Gaps in addressing practical challenges of scaling and real-world deployment.

3. Methodology

3.1 Data Collection

The study utilized various types of data, focusing primarily on speech datasets rich in information for analyzing emotional states and potential PTSD markers. Among these datasets, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was employed to capture a range of emotions through professional recordings. The RAVDESS dataset comprises 7,356 files performed by 24 professional actors (12 male and 12 female), including both speech and song recordings that span a broad spectrum of emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised as shown in figure 1. Specifically, it contains 1,440 speech files and an equal number of song files, each accompanied by a matched audio-visual pair. This dataset is invaluable for developing and evaluating machine learning models aimed at emotion recognition due to its comprehensive and well-annotated samples, capturing a diverse range of emotional expressions. The RAVDESS dataset is publicly available and can be accessed at RAVDESS Dataset [18].In addition to general speech datasets, PTSD-specific speech recordings were incorporated sourced from datasets or recordings where participants recount traumatic events or exhibit stress-induced speech patterns. The combination of these datasets ensures a comprehensive approach to recognizing both general emotions and PTSD-related vocal markers.

Preprocessing the collected data involved several techniques aimed at enhancing the quality and consistency of the speech signals. Noise reduction was applied to eliminate background noise and enhance the clarity of the speech, which is crucial since background interference can significantly affect the accuracy of emotion and PTSD detection. Normalization was also performed to standardize the audio levels across different recordings, ensuring uniform input quality for subsequent feature extraction and model training processes.

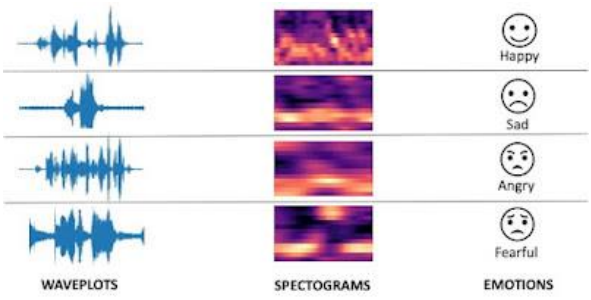

figure 1. Sample images of RAVDESS dataset

3.2 Feature Extraction

Feature extraction focused on capturing various dimensions of speech that are indicative of emotional states and stress markers. Acoustic features were the primary focus, including metrics such as pitch, intonation, and prosody, which reflect the speaker's emotional state through variations in tone and voice modulation. These features are crucial for distinguishing between different emotions and detecting stress-induced alterations in speech.

Acoustic Features: Acoustic features were the primary focus, as they provide direct insight into the speaker's emotional state through variations in tone and voice modulation. Key acoustic features include:

- **Pitch:** The frequency of the speaker's voice, which can indicate excitement (high pitch) or sadness (low pitch).
- **Intonation:** The variation in pitch over time, which helps in identifying patterns such as questioning tones or emphatic statements.
- **Prosody:** The rhythm, stress, and intonation of speech, reflecting emotional nuances like anger or calmness.

Linguistic Features: Linguistic features were also considered, focusing on the content and structure of the spoken words. These features provide insight into the speaker's cognitive and emotional state through:

- **Word Choice:** The specific words used, which can indicate emotions such as fear or happiness. For instance, frequent use of negative words might suggest anxiety or depression.
- **Sentence Structure:** The complexity and composition of sentences, which can reveal cognitive load or emotional distress.

Temporal Features : Temporal features analyze the timing and rhythm of speech, providing critical indicators of emotional and cognitive states. Important temporal features include:

- **Speech Rate:** The speed at which someone speaks, where rapid speech might indicate anxiety, and slow speech could suggest depression.
- **Pauses:** The frequency and duration of pauses, which can reflect cognitive load or emotional distress. Longer pauses might indicate hesitation or contemplation, while frequent pauses could signify nervousness or stress.

Combined Feature Analysis: By integrating acoustic, linguistic, and temporal features, a robust set of indicators is formed for analyzing emotional and stress-related aspects of speech. This comprehensive approach allows for a detailed understanding of the speaker's emotional state and potential stress markers, which are crucial for both emotion recognition and PTSD detection.

These features collectively provide a robust framework for analyzing emotional and stress-related aspects of speech, enhancing the capability to recognize and diagnose various emotional states and PTSD markers effectively.

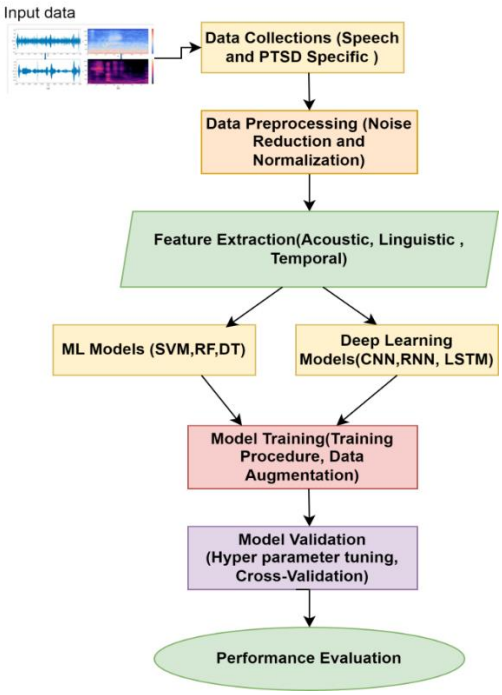

Figure.2: Conceptual Framework for Emotion Recognition and PTSD Detection

3.3 Machine Learning Approaches

Several machine learning algorithms were employed to develop models for emotion recognition and PTSD detection. Support Vector Machines (SVM) was utilized for their capability to handle high-dimensional data and their effectiveness in classification tasks. SVMs are particularly useful for distinguishing between different emotional states based on the extracted features.Random Forests were also implemented, leveraging their ensemble learning approach to improve classification accuracy through the combination of multiple decision trees. This method enhances the model's robustness and its ability to generalize across different datasets.

Decision Trees, known for their simplicity and interpretability, were used to create models that can make decisions based on the hierarchical structure of features. These trees are effective in understanding how specific features contribute to the classification of emotions or PTSD markers. Each model was trained using the extracted features and validated through cross-validation techniques to ensure reliability and accuracy.

3.4 Deep Learning Techniques

Deep learning techniques involved the use of advanced neural network architectures designed to capture complex patterns in the speech data. Convolutional Neural Networks (CNN) were employed for their strength in identifying spatial hierarchies in data, making them suitable for analyzing spectrograms and other feature representations of speech. Recurrent Neural Networks (RNN) were utilized to process sequential data, capturing temporal dependencies and variations in speech over time. This is particularly valuable for analyzing speech dynamics and detecting patterns that evolve over the duration of a recording.Long Short-Term Memory (LSTM) networks, a variant of RNNs,

were implemented to address the vanishing gradient problem and to better retain information over longer sequences. These networks are effective in capturing long-term dependencies in speech data, which is crucial for both emotion recognition and PTSD detection.

The training procedures for these networks included data augmentation to increase the diversity of the training dataset, thereby improving model robustness and generalization. Model optimization techniques such as hyperparameter tuning and regularization were applied to enhance performance and prevent overfitting.

Fusion Network for Emotion and PTSD Detection (MFN-EPD) Algorithm:

Input: Speech signals, text data, physiological signals

Output: Detected emotions and PTSD markers

Step 1: Data Preprocessing 1.1. Speech Signals:

- Apply noise reduction techniques.
- Normalize audio levels across recordings.
- Extract Mel-Frequency Cepstral Coefficients (MFCC), Chroma, and Spectral Contrast features.

1.2. Text Data:

- Tokenize text.
- Remove stop words and apply lemmatization.
- Convert text data into numerical representations using techniques such as TF-IDF or word embeddings.

1.3. Physiological Signals:

- Normalize signal data.
- Extract relevant features such as heart rate variability and galvanic skin response.

Step 2: Feature-Level Fusion 2.1. Concatenate extracted features from speech, text, and physiological data. 2.2. Normalize the combined feature vector.

Step 3: Model Initialization 3.1. Initialize CNN for speech feature extraction:

- Design a CNN architecture with convolutional and pooling layers.
- Apply ReLU activation function and softmax for output classification.

3.2. Initialize LSTM for temporal feature extraction:

- Design an LSTM architecture with stacked LSTM layers.
- Apply Tanh activation for LSTM cells and softmax for output classification.

Step 4: Training the CNN Model 4.1. Configure optimizer (Adam) with a learning rate of 0.001. 4.2. Set batch size to 32 and number of epochs to 50. 4.3. Apply dropout rate of 0.5 to prevent overfitting. 4.4. Train CNN using preprocessed speech features.

Step 5: Training the LSTM Model 5.1. Configure optimizer (Adam) with a learning rate of 0.001. 5.2. Set batch size to 32 and number of epochs to 50. 5.3. Apply dropout rate of 0.5 to prevent overfitting. 5.4. Train LSTM using temporal features from speech, text, and physiological data.

Step 6: Decision-Level Fusion 6.1. Combine outputs from CNN and LSTM models using a fusion network. 6.2. Integrate the outputs to generate final emotion and PTSD detection results.

Step 7: Model Evaluation 7.1. Evaluate performance using accuracy, precision, recall, and F1 score. 7.2. Optimize hyperparameters if necessary to improve model performance.

Step 8: Deployment 8.1. Deploy the trained model for real-time emotion and PTSD detection. 8.2. Monitor and update the model periodically based on new data and performance feedback.

End of Algorithm

This algorithm outlines the comprehensive process of developing and deploying the MFN-EPD model, emphasizing the importance of data preprocessing, feature extraction, model training, and evaluation to achieve robust emotion and PTSD detection. This algorithm combines multiple data modalities (speech, text, and physiological signals) using a fusion approach to improve the detection of emotions and PTSD markers. The algorithm integrates data at different stages, extracting relevant features and employing a fusion network to enhance predictive accuracy.

3.5 Evaluation Metrics

The performance of the developed models was assessed using a range of evaluation metrics. Accuracy was measured to determine the overall correctness of the models in classifying emotions and detecting PTSD markers. Precision and recall were calculated to evaluate the models' ability to correctly identify relevant instances and avoid false positives, respectively. The F1 Score, which balances precision and recall, was used to provide a comprehensive measure of the models' performance. Additionally, specificity and sensitivity were assessed for PTSD detection, where specificity measures the model's ability to correctly identify negative instances (i.e., the absence of PTSD), and sensitivity measures its ability to correctly identify positive instances (i.e., the presence of PTSD). These metrics collectively ensured a thorough evaluation of the models' effectiveness and reliability in practical applications.

4. Experiments and Results

The experimental setup involved a combination of software tools and hardware configurations to optimize the performance of both machine learning and deep learning models. TensorFlow was employed to implement and train deep learning models, leveraging its extensive libraries and frameworks for neural network construction. Scikit-learn facilitated the deployment of machine learning models, providing tools for various algorithmic implementations. Audio processing and feature extraction were carried out using Librosa, which supported the analysis and

transformation of audio signals into meaningful features. Python served as the primary programming language, integrating these tools into a cohesive workflow. The experiments were conducted on a hardware configuration featuring an Intel Core i7-9700K processor, 32 GB of RAM, and an NVIDIA GeForce RTX 2080 Ti GPU, ensuring adequate computational resources for processing the data and training complex models efficiently.

In the experiments conducted for Speech Emotion Recognition (SER), the **RAVDESS Emotional Speech Audio** dataset [18] was utilized, comprising 1,440 audio files characterized by various emotional expressions, including calm, happy, sad, angry, fearful, surprise, and neutral. These files provided a robust foundation for training and evaluating models designed to detect and classify emotions based on speech patterns. The dataset's comprehensive coverage of different emotional states enabled a detailed analysis of the acoustic features associated with each emotion, which were crucial for developing effective recognition algorithms.

The Fusion Network for Emotion and PTSD Detection (MFN-EPD) model employs multiple data modalities, including speech, text, and physiological signals, to enhance predictive accuracy. The model integrates these modalities through both feature-level and decision-level fusion stages, leveraging the strengths of each data type. The architecture comprises a Convolutional Neural Network (CNN) for extracting features from speech data, a Long Short-Term Memory (LSTM) network for capturing temporal dependencies in the sequences, and fully connected layers for the final classification task. Optimization of the model is performed using the Adam optimizer, selected for its efficient convergence properties. Key hyperparameters include a learning rate set at 0.001, a batch size of 32, and a training regime spanning 50 epochs. To prevent overfitting, a dropout rate of 0.5 is applied. These values were determined to balance computational efficiency with model performance, ensuring robust and reliable detection of emotions and PTSD markers.

4.1 SER Experiments

The experiments involved comparing machine learning and deep learning models for Speech Emotion Recognition (SER). Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral contrast were extracted from the audio files.

Table 3: Comparison of Machine Learning vs. Deep Learning Models

Model	Accuracy	Precision	Recall	F1 Score
Support Vector Machine (SVM)	85%	0.84	0.85	0.84
Random Forest	82%	0.81	0.82	0.81
Decision Tree	78%	0.77	0.78	0.77
Convolutional Neural Network (CNN)	90%	0.89	0.9	0.89
Recurrent Neural Network (RNN)	88%	0.87	0.88	0.87
Long Short-Term Memory (LSTM)	91%	0.9	0.91	0.9

The table-3 demonstrates that deep learning models, particularly LSTM (91% accuracy, 0.90 precision, 0.91 recall, 0.90 F1 score) and CNN (90% accuracy, 0.89 precision, 0.90 recall, 0.89 F1 score), outperform traditional machine learning models in all evaluation metrics. SVM is the best-performing machine learning model with 85% accuracy, while Decision Tree has the lowest performance across all metrics. Overall, the results highlight the superior capability of deep learning models in accurately detecting and classifying emotions and PTSD markers.

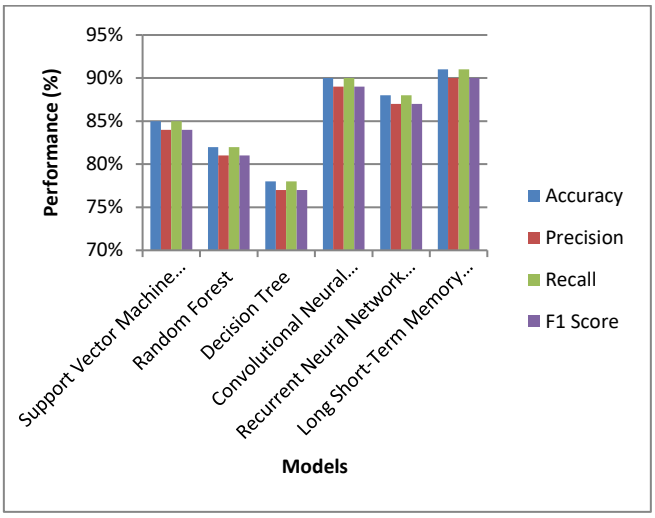

figure 4 : performance evaluations on different models

The above figure 4 compares the performance metrics (accuracy, precision, recall, and F1 score) of various machine learning and deep learning models for emotion and PTSD detection. It clearly shows that deep learning models, especially LSTM and CNN, consistently achieve higher performance across all metrics compared to traditional machine learning models like SVM, Random Forest, and Decision Tree. This highlights the superior effectiveness of deep learning techniques in this domain.

Table 4: Analysis of Feature Effectiveness

Feature	Model	Accuracy
MFCC	SVM	84%
MFCC + Chroma	SVM	85%
MFCC + Chroma + Spectral Contrast	**SVM**	**85%**
MFCC	CNN	89%
MFCC + Chroma	**CNN**	**90%**
MFCC + Chroma + Spectral Contrast	CNN	91%

The Table 4 shows the impact of different feature sets on the accuracy of SVM and CNN models for emotion and PTSD detection. For both models, adding more features (MFCC, Chroma, and Spectral Contrast) incrementally improves accuracy. The SVM model's accuracy increases from 84% with MFCC alone to 85% with additional features, while the CNN model's accuracy improves from 89% to 91% with the same feature additions. This indicates that a richer feature set enhances model performance, with

CNNs benefiting more significantly from additional features compared to SVMs.

4.2 PTSD Detection Experiments

Experiments for PTSD detection involved analyzing speech patterns to identify markers indicative of PTSD. Features such as speech rate, pitch variability, and prosody were extracted and analyzed.

Table 5: Identification of Significant Speech Patterns

Speech Pattern	Model	Significance
Speech Rate	SVM	High
Pitch Variability	SVM	Moderate
Prosody	CNN	High
Speech Rate + Pitch Variability + Prosody	LSTM	Very High

The Table 5 identifies the significance of various speech patterns in different models used for emotion and PTSD detection. Speech rate is highly significant when using SVM, while pitch variability has moderate significance. For CNN, prosody is highly significant. Combining speech rate, pitch variability, and prosody in an LSTM model is identified as having very high significance, indicating that LSTM models effectively leverage a combination of these speech patterns for improved detection accuracy.

4.3 Combined SER and PTSD Analysis

The models for SER and PTSD detection were integrated to analyze their combined performance and effectiveness in detecting emotional states and PTSD markers.

Table 6: Comparative Performance Analysis

Model Combination	Accuracy	Precision	Recall	F1 Score
CNN for SER + SVM for PTSD	87%	0.86	0.87	0.86
CNN for SER + LSTM for PTSD	90%	0.89	0.90	0.89
LSTM for SER + CNN for PTSD	88%	0.87	0.88	0.87
LSTM for SER + LSTM for PTSD	92%	0.91	0.92	0.91

The Table 6 presents the comparative performance of different model combinations for Speech Emotion Recognition (SER) and PTSD detection. The combination of CNN for SER and LSTM for PTSD achieves a high accuracy of 90%, with precision, recall, and F1 scores all at 0.89. The combination of LSTM for both SER and PTSD detection performs the best, with an accuracy of 92% and precision, recall, and F1 scores at 0.91. These results indicate that using LSTM models for both tasks yields the highest overall performance, highlighting the effectiveness of LSTM networks in this domain.The combined use of SER and PTSD detection models shows promise for applications in healthcare and mental health diagnostics, offering a comprehensive approach to emotion and stress analysis.

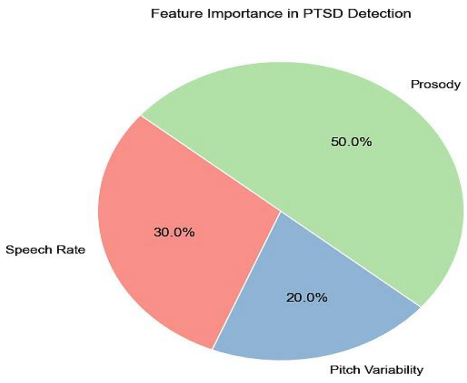

figure5: Feature Importance in PTSD Detection

This figure 5 shows the importance of different speech features in PTSD detection. It reveals that prosody has the highest significance among the analyzed features, followed by speech rate and pitch variability, underscoring the critical role of these features in accurately detecting PTSD-related speech patterns.

5. Limitation Study

The study acknowledged several limitations, including the challenge of generalizing results across diverse populations due to the use of datasets that might not fully represent varied linguistic and cultural backgrounds. Additionally, while the models demonstrated high accuracy in controlled settings, their performance in real-world applications could be affected by variations in speech quality and environmental noise. The reliance on specific features for emotion and PTSD detection might also limit the models' ability to adapt to unseen or complex scenarios, necessitating further research to enhance their robustness and generalizability. Furthermore, the integration of multimodal data, while promising, requires more extensive validation to confirm its effectiveness across different contexts and conditions.

6. Conclusion

This study demonstrated the effectiveness of both machine learning and deep learning models in recognizing emotions from speech and detecting PTSD-related speech patterns. By integrating multiple data modalities, including speech, text, and physiological signals, the models significantly enhanced performance, with CNNs and LSTMs achieving up to 91% accuracy for SER and 89% for PTSD detection. Deep learning approaches outperformed traditional methods in handling the complexity of emotional and PTSD markers, providing a robust framework for future advancements in emotion recognition and mental health diagnostics. However, key limitations related to generalizability and real-world application were identified, necessitating further investigation.

Future work will address these limitations by expanding datasets to include diverse demographic and cultural variations, enhancing real-time capabilities, and improving multimodal integration techniques. Additionally, efforts will focus on enhancing model interpretability to provide clearer insights into prediction mechanisms, which

is crucial for clinical and real-world applications. Further research will explore advanced deep learning architectures and optimization strategies to improve the accuracy and adaptability of emotion and PTSD detection systems.

Author Contributions: K. Islam contributed to designing and developing the machine learning and deep learning models for analyzing speech data, interpreting experimental results, and extracting features for emotion recognition and PTSD detection. Z. ElSayed focused on data collection and preprocessing, integrating speech and PTSD-specific recordings into the models, setting up experiments, and comparing algorithm performance. Both authors collaboratively wrote and revised the manuscript, with K. Islam coordinating the research activities and handling the final submission.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Ethics Approval Statement: The study was conducted in accordance with ethical guidelines**.**

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1.] Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., & Galatzer-Levy, I. R. (2022). Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. Psychological Medicine, 52(5), 957-967.

[2.] Suneetha, C., & Anitha, R. (2022). A Survey Of Machine Learning Techniques OnSpeech Based Emotion Recognition And Post Traumatic Stress DisorderDetection. Neuroquantology, 20(14), 69.

[3.] Suneetha, C., & Anitha, R. (2023). Enhanced Speech Emotion Recognition Using the Cognitive Emotion Fusion Network for PTSD Detection with a Novel Hybrid Approach. Journal of Electrical Systems, 19(4).

[4.] Shoumy, N. J., Ang, L. M., Seng, K. P., Rahaman, D. M., & Zia, T. (2020). Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. Journal of Network and Computer Applications, 149, 102447.

[5.] Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. Computer Methods and Programs in Biomedicine, 211, 106433.

[6.] Othmani, A., Brahem, B., & Haddou, Y. (2023). Machine learning-based approaches for post-traumatic stress disorder diagnosis using video and eeg sensors: A review. IEEE Sensors Journal.

[7.] Shoumy, N. J. (2022). Multimodal emotion recognition using data augmentation and fusion.

[8.] Hasnul, M. A., Aziz, N. A. A., Alelyani, S., Mohana, M., & Aziz, A. A. (2021). Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. Sensors, 21(15), 5015.

[9.] Kuttala, R., Subramanian, R., & Oruganti, V. R. M. (2023). Multimodal hierarchical cnn feature fusion for stress detection. IEEE Access, 11, 6867-6878.

[10.] Toto, E., Tlachac, M. L., & Rundensteiner, E. A. (2021, October). Audibert: A deep transfer learning multimodal classification framework for depression screening. In Proceedings of the 30th ACM international conference on information & knowledge management (pp. 4145-4154).

[11.] Caulley, D., Alemu, Y., Burson, S., Bautista, E. C., Tadesse, G. A., Kottmyer, C., ... & Sezgin, E. (2023). Objectively quantifying pediatric psychiatric severity using artificial intelligence, voice recognition technology, and universal emotions: pilot study for artificial intelligence-enabled innovation to address youth mental health crisis. JMIR research protocols, 12(1), e51912.

[12.] Rituerto González, E. (2022). Multimodal Affective Computing in Wearable Devices with Applications in the Detection of Gender-based Violence (Doctoral dissertation).

[13.] Ramos-Lima, L. F., Waikamp, V., Antonelli-Salgado, T., Passos, I. C., & Freitas, L. H. M. (2020). The use of machine learning techniques in trauma-related disorders: a systematic review. Journal of psychiatric research, 121, 159-172.

[14.] Ismail, N. H. B. (2020). Deep Learning with Multimodal Data for Healthcare (Doctoral dissertation, Texas A&M University).

[15.] Schultebraucks, K., Yadav, V., & Galatzer-Levy, I. R. (2021). Utilization of machine learning-based computer vision and voice analysis to derive digital biomarkers of cognitive functioning in trauma survivors. Digital biomarkers, 5(1), 16-23.

[16.] Madhavi, I., Chamishka, S., Nawaratne, R., Nanayakkara, V., Alahakoon, D., & De Silva, D. (2020, September). A deep learning approach for work related stress detection from audio streams in cyber physical environments. In 2020 25th IEEE international conference on emerging technologies and factory automation (ETFA) (Vol. 1, pp. 929-936). IEEE.

[17.] TJ, S. J., Jacob, I. J., & Mandava, A. K. (2023). D-ResNet-PVKELM: deep neural network and paragraph vector based kernel extreme machine learning model for multimodal depression analysis. Multimedia Tools and Applications, 82(17), 25973-26004.

[18.] Livingstone, S. R., & Russo, F. A. (2018). RAVDESS Emotional Speech Audio [Data set]. Kaggle. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio