**Multi-Modal Audio Feature Extraction and Dual CNN Fusion for PTSD Severity Assessment**

Arjita Choubey1, a), Manoj Kumar Pandey1, b), Ashwani Kumar Dubey1, c), Angela Amphawan2, d) and Tse-Kian Neo3, 4, e)

*1Department of Electronics and Communication Engineering, ASET, Amity University Uttar Pradesh, Noida India.*

*2Smart Photonics Lab, Sunway University, Malaysia.*

*3Faculty of Creative Multimedia, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia.*

*4Centre for Innovative and Immersive Technology, CoE for Immersive Experience, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia.*

*e) Corresponding author: tkneo@mmu.edu.my*

*a) arjitaece1015@gmail.com*

*b) mkpandey1@amity.edu*

*c) dubey1ak@gmail.com*

*d) angelaa@sunway.edu.my*

**Abstract.** Post-traumatic stress disorder is a disorder identified as the inability to bounce back from a traumatic experience. For assessingthe severity of this disorder using speech, a novel framework is presented in this paper. This method is proposed over an extended- DAIC dataset. It uses multi-modal feature extraction that is, spectrograms, MFCCs, and Mel-spectrograms. These features are processed by a Dual-CNN model that fuses outputs of two CNN branches with different features. The feature selection uses a two-level approach that includes ANOVA and ABC optimizer for refinement. Linear kernel SVM classifies the features into four clinically relevant categories. This framework effectively processes the various representations and further selection of features results in an accuracy of 98.4%.

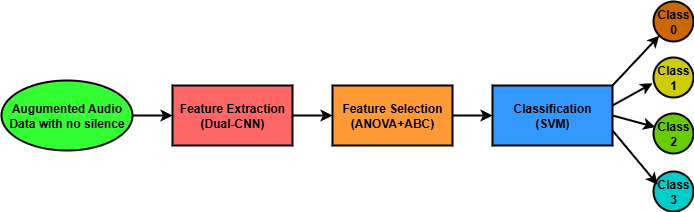
# INTRODUCTION

Experiencing or witnessing a traumatic event always leaves a scar on people’s hearts. Traumas and people’s reactions to the trauma reduce with time. When these reactions persist for an extended duration and begin to disturb their day-to-day life then this situation might be classified as Post-traumatic stress disorder (PTSD). As per WHO, approximately 70% of people in the world go through traumatic experiences during their lifespan. Approximately 3.9% population has sometimes faced PTSD in their life [1].

For detection of the PTSD and its severity various scoring methods are used. PCL-5 and PCL-C are two self-report questionnaires in which a person ratesthe severity of different PTSD symptoms on a scale. The higher the PCL-C score, higheris the PTSDseverity. PCL-5 is for veterans and PCL-C is for Civilians. In this paper, PCL-C has been taken for labeling audio files in four classes. Scores 17-20 represent no to minimum symptoms of PTSD (class 0), scores 21-29 represent mild symptoms of PTSD (class 1), scores 30-49 represent moderate symptoms of PTSD (class 2) and scores 50-86 represent severe symptoms of PTSD (class 3) [2]. Using speech for the detection of the severity of PTSD is still an open area of research. By using this scoring method for labeling and utilizing different feature extraction techniques for properly capturing the acoustic characteristics of speech such as spectrograms, Mel-spectrograms, MFCCsetc, a framework can be made for assessing the severity of PTSD using speech whichwill be beneficial for both patients and doctors for timely detection and quick assessment of PTSD.While some work has been done on speech based binary classification of PTSD but PTSD assessment for detecting the severity levels is still under-explored. The presented work provides a novel approach for multi-class (severity) classification of PTSD and achieves accuracy of 98.4% and thus contributes in the emerging field of multi-class classification of PTSD as it can be integrated into telehealth platforms or mobile applications for early detection and monitoring.

In this paper, a novel approach is presented that detects the severity levels of PTSD with a good accuracy of 98.4%. Figure 1 shows the framework that includes:

1. Feature extraction: A novel Dual-CNN feature extractor that consists of two branches. One branch passes the spectrograms through Glorot initialized CNN model and the other branch passes the Mel-spectrogram and MFCC features through the identical CNN model used in the first branch.
2. Feature selection: A two-level feature selector that uses ANOVA (for p< 0.05) and then ABC optimizer for refinement.
3. Multi-level classification: SVM with Linear kernel is used for classification.



**FIGURE 1.** Block diagram of the proposed framework with novel Dual-CNN model for feature extraction

# RELATED WORK

In this paper, interview data from 148 individuals who have faced the same terrorist attack were collected. This data was then analyzed to perform binary classification under three domains that are psychiatry, linguistics, and the Natural Language Processing (NLP) community. For the first domain, psychiatry-only transcription was used and the result with an AUC of 0.72 is achieved. For the linguistics domain, statistical analysis and machine learning models were used to achieve an AUC of 0.69, and for the last domain i.e. Natural Language Processing 0.64 AUC was obtained. This work provides ground for a relationship between PTSD and language [3].

In this paper, EMODB and RAVDESS datasets were used to get an accuracy of 98.11% and 91.17% respectively. Behind this result crucial contribution is the proposed Cognitive Emotion Fusion Network (CEFNet). It is a hybrid model that integrates improved and faster Region-based Convolutional Neural Networks (IFR-CNN), Deep Convolutional Neural Networks (DCNNs), Deep Belief Networks (DBNs), and the Bird's Nest Learning Analogy (BNLA) and performs better as compared to the existing models [4].

In this paper, an approach is proposed that follows three steps: pre-processing, feature extraction, and binary classification. Pre-processing is done by segmenting the speech and feature extraction and classification are performed by the XGBoost-based Teamwork optimization (XGB-TWO) algorithm. Two datasets, TIMIT and FEMH are used to obtain an accuracy of 98.25% [5].

This paper proposes a novel Multi-Strategy Seeker Archimedes Optimization-based Elman Recurrent Neural Network (MSSAO-ERNN) for the binary classification of PTSD. This model includes all the stages including the pre-processing stage, feature extraction stage, and classification stage. The model is implemented on three datasets: the NNE dataset, the FME hospital dataset, and the TIMIT dataset, and achieves an accuracy of 97% [6].

In this paper, a dataset is recorded that consists of 76 PTSD patients and 60 healthy individuals. The feature extraction openSmile framework is used here and feature selection is done by random forest algorithm. After selection, six different classification models and a regression model are used to compare different results. The highest accuracy (0.975) is achieved by the RF model with the highest AUC of 1.0. In the case of a regression model, the result achieved is: MSE=0.90, MAE=0.76, , p<0.01[7].

The focus of majority of works done in PTSD is binary classification (PTSD vs. non-PTSD) and thus restricted comprehension about the severity level of PTSD is provided. The presented work provides better results as compared to the only work [7] done in multi-class classification (severity) of PTSD. Thus, contributing to the multi-class classification (severity) of PTSD which has clinical relevance as understanding the severity is important for customized support for this mental disorder.

# METHODOLOGY

## Data Preparation

The Extended Distress Analysis Interview Corpus (E-DAIC) dataset [8] is used for the proposed work, which is an extended version of WOZ-DAIC. This dataset consists of semi-clinical interviews to support the assessment of mental disorders by a virtual interviewer. This dataset is labeled for PTSD severity using the PCL-C method. The data preparation stage consists of three parts:

1. Removal of silence and voice of the interviewer [9].
2. Since the dataset is imbalanced, the time-shifting augmentation technique is applied and 80 audio data is taken for each class. This creates a balanced database of 320 audio files. The dataset will be processed further concerning Participant IDs and associated PTSD Severity scores.
3. Labeling of audio files in terms of four classes. Class 0: Scores 17-20, Class 1: Scores 21-29, Class 2: Scores 30-49 and Class 3: Scores 50-86. A unique participant ID is assigned to all the augmented data.

## Feature Extraction

For the feature extraction stage, three features are extracted in this paper: Spectrogram, Mel-spectrogram, and MFCC. The spectrogram is computed using the Short-Time Fourier Transform (STFT) and the amplitude is converted to dB scale.Mel-spectrogram is extractedand the amplitude is converted to dB scale.A Mel-spectrogram is obtained by mapping the power spectrogram (which is typically computed using an STFT) onto the Mel-scale using a bank of filters. The Mel-scale frequency mapping is given by Equation (1)[10]:

(1)

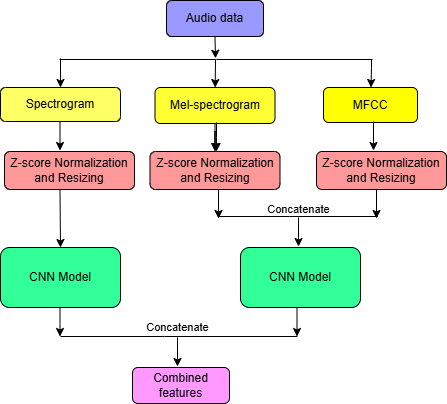
Where, is the perceived frequency and is the real linear frequency in speech signal.

And MFCC (Mel Frequency Cepstral Coefficients) are extracted. For MFCC calculation the following pipeline is followed [11]. The continuous speech signal is passed through a window then DCT is applied followed by Mel-frequency wrapping. Log is calculated for the generated Mel-spectrum and finally, inverse DFT produces Mel-cepstrum.

All the features are then passed through processes of Z-score normalization and resizing. Z-score normalization normalizes a feature array by subtracting the mean and dividing by the standard deviation, with a small constant added to avoid division by zero. This transformation rescales the data so that it has a mean of 0 and a standard deviation of 1, which is especially useful when features have different units or scales.

The feature extractor now introduces a Dual-CNN model of identical characteristics. A spectrogram is passed through one CNN and a second CNN concatenated Mel-spectrogram and MFCC is passed. The CNN models are designed to process 2D inputs that are spectrogram images and combine Mel spectrogram and MFCC features. They consist of three convolutional block layers, succeeded by a flattening layer to convert the multi-dimensional output into a 1D vector for further processing or classification. They consist of 3 x 3 kernels with Glorot initialization to ensure weight initialization in a balanced way and LeakyReLU Activation with alpha = 0.01. The number of features extracted by this process is 268800. Removal of irrelevant or redundant features will be done in the next stage which is feature selection.

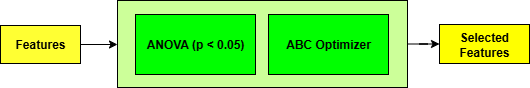
The dual CNN is designed to process different representations of the audio signal separately (see Figure 2). While Mel+MFCC capture most of the discriminative information, the spectrogram branch is intended to capture additional, potentially complementary, time-frequency patterns. By fusing outputs from two specialized networks, the dual CNN framework offers enhanced robustness. In more challenging or noisy environments the additional branch will help generalize better, reducing overfitting to specific features that could be present in only one modality.

**FIGURE 2.** Block diagram feature extractor with Dual-CNN

## Feature Selection

Feature selection helps make models simpler, faster, and often more accurate, all while making the results more interpretable and robust. For this stage, a two-level approach is used as shown in Figure 3. Level 1 is the Analysis of Variance (ANOVA) which acts as a statistical filter and identifies features that are statistically important in differentiating among the classes. The filtration process is done by computing the F-statistic and corresponding p-values. Features that have p-values below the predefined threshold (p< 0.05) were selected as such features demonstrated a strong statistical relation with class labels. Level 2 is an Artificial Bee Colony (ABC) optimizer that refines the features selected by ANOVA [12]. The Artificial Bee Colony (ABC) algorithm is a swarm intelligence-based optimization technique inspired by how honey bees identify suitable nectar and pollen near them.

This two-level feature selector significantly reduces the number of features, keeping only the most significant ones. This ensures a better case scenario for the upcoming stage of multi-level classification of PTSD.



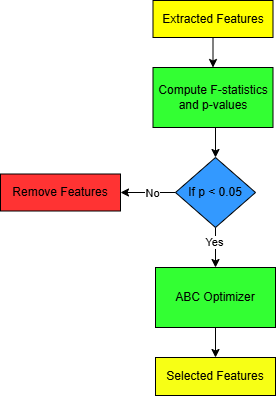
**FIGURE 3.** Two-level feature selector

Figure 4 shows the diagram of a two-level feature selector. The F-statistics and p-value are computed for extracted features. If features have p > 0.05 then those features are removed and only features with p < 0.05 are passed to the next level for refinement using ABC optimizer. The feature selected by the ABC optimizer will be the final selected feature.

# Multi-level Classification

For the Multi-level classification of PTSD based on the features selected by the above stage, a Support Vector Machine (SVM) is used. SVM targets to find the optimal hyperplane that maximizes the margin between the two classes. A kernel enables SVMs to perform non-linear classification by implicitly mapping the input features into a high-dimensional feature space. The kernel used in the proposed work was Linear. It handles high-dimensional data effectively and provides robustness to overfitting and hence yields a good result of 98.4%.

To make sure that all features contribute equally to the learning process of the model, stratified sampling is ensured and the dataset was divided into training (64%), validation (16%), and test (20%) subsets. A training set is used for training, a validation set is used for tuning the hyperparameter, and a test set is used to evaluate the accomplishments of the model.

**FIGURE 4.** Block diagram of two-level feature selector

# RESULTS AND DISCUSSION

The framework presented in the paper provides 98.4% accuracy in the detection of severity levels of PTSD. Following are all the metrics used for understanding the effectiveness of the work. Table 1 shows class-wise Precision, recall, F1 Score, and accuracy of classification that are calculated by using Equations (2), (3), (4) and (5) respectively. Here, Classes 0 and 2 have perfect Precision, Recall, F1 Score and Accuracy. One instance from class 3 was “mistakenly” predicted as 1 and Class 3 missed one sample, so class 1’s precision and class 3’s 0.94 recall is 0.94. The average accuracy is 98.4%, which shows how well classification is done, however need for a larger dataset for more validation remains. Figure 5 shows the confusion matrix. A confusion matrix compares the true class labels (rows) to the predicted labels (columns). The diagonal cells are the correctly predicted samples. Off‐diagonal cells are misclassifications.

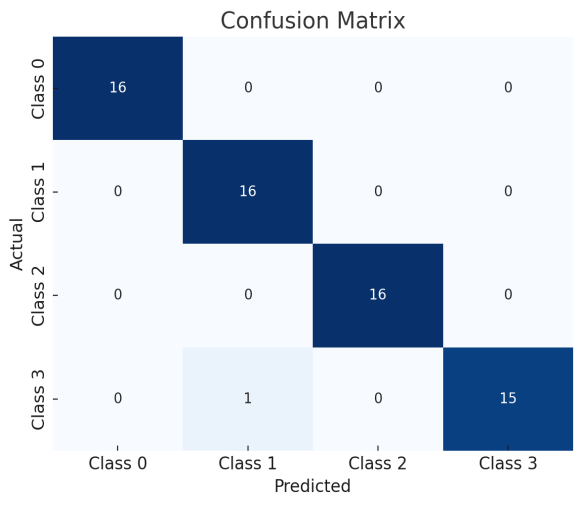
(2)

(3)

(4)

(5)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TABLE 1.** Classification report of framework showing precision, recall, F1 score, and accuracy for each class | | | | |
| **Class** | **Precision** | **Recall** | **F1 Score** | **Accuracy** |
| 0 | 100% | 100% | 100% | 100% |
| 1 | 94% | 100% | 97% | 100% |
| 2 | 100% | 100% | 100% | 100% |
| 3 | 100% | 94% | 97% | 93.75% |



**FIGURE 5.** Confusion matrix showing only one misclassification

The other matrices that are calculated to see the efficacy of the model are Mean Squared Error (MSE) and Mean Absolute Error (MAE), calculated by Equation (6) and Equation (7), respectively.

MSE = (6)

MAE = (7)

Concerning the comparison of the work with previous work, only one work has been done yet that provides a multi-class classification for PTSD. This work uses openSmile for the feature extraction process and Random forest is used for classification. However, the data used is different but it is a smaller dataset as compared to our dataset [7]. The dataset used in [7] contains only 136 participants (76 PTSD patients and 60 healthy individuals). However, the dataset used in the presented work contains 320 audio files. This dataset is created by carefully balancing the E-DAIC dataset by using augmentation technique. Precision, recall, F1 Score, accuracy, MSE (p<0.001), and MAE (p<0.001)are calculated in the paper as shown in Table 2 along with matrices of the proposed work. The output of the proposed work is classification labels (Class 0 to Class 3), therefore the appropriate metrics are precision, recall, F1-score, and accuracy. The regression matrices (MSE and MAE) are included to provide comparative reference with the study by J. Hu et al. [7], which includes both classification and regression tasks.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TABLE 2.** Comparison of multi-level PTSD detection works | | | | | | |
| **Paper** | **Precision** | **Recall** | **F1 Score** | **Accuracy** | **MSE** | **MAE** |
| J. Hu. et al [7] | 97.3% | 97.1% | 97.1% | 97.5% | 0.90 | 0.76 |
| Proposed work | 99% | 98% | 98% | 98.4% | 0.0635 | 0.0312 |

# CONCLUSION

This paper puts forward a complete framework for the detection of PTSD severity with a novel Dual-CNN feature extractor technique that uses features from spectrogram, Mel-spectrogram, and MFCC. The work provides great success in the domain of lesser-explored PTSD severity detection and provides an accuracy of 98.4%. Feature selection is done by a two-level approach of ANOVA and ABC optimizer and SVM classifier that performs effectively while handling the high-dimensional data. This work is a step forward in the detection of the severity of PTSD and presents a notable tool for early assessment and for providing customized strategies for the management of PTSD.

# ACKNOWLEDGEMENT AND FUNDING

We would like to acknowledge the support of TM R&D Grant (RDTC/231106; MMUE/230053) from TM R&D Sdn. Bhd., MMU and thank the members for their contribution and collaboration towards this research publication.

# REFERENCES

1. R.C. Kessler, S. Aguilar-Gaxiola, J. Alonso, C. Benjet, E.J. Bromet, G. Cardoso, L. Degenhardt, G. De Girolamo, R.V. Dinolova, F. Ferry, S. Florescu, O. Gureje, J.M. Haro, Y. Huang, E.G. Karam, N. Kawakami, S. Lee, J.-P. Lepine, D. Levinson, F. Navarro-Mateu, B.-E. Pennell, M. Piazza, J. Posada-Villa, K.M. Scott, D.J. Stein, M. Ten Have, Y. Torres, M.C. Viana, M.V. Petukhova, N.A. Sampson, A.M. Zaslavsky, and K.C. Koenen, “Trauma and PTSD in the WHO World Mental Health Surveys,” European Journal of Psychotraumatology **8**(sup5), 1353383 (2017).
2. F.W. Weathers, B.T. Litz, J.A. Huska, and T.M. Keane, PTSD Checklist – Civilian Version (PCL-C) (National Center for PTSD, Behavioral Science Division, Boston, MA, 1994).
3. R. Quillivic, F. Gayraud, Y. Auxéméry, L. Vanni, D. Peschanski, F. Eustache, J. Dayan, and S. Mesmoudi, “Interdisciplinary approach to identify language markers for post-traumatic stress disorder using machine learning and deep learning,” Scientific Reports **14**(1), 12468 (2024).
4. E.Al. Chappidi Suneetha, “Enhanced Speech Emotion Recognition Using the Cognitive Emotion Fusion Network for PTSD Detection with a Novel Hybrid Approach,” Journal of Engineering Science **19**(4), 376–398 (2024).
5. S. Gupta, L. Goel, A. Singh, A.K. Agarwal, and R.K. Singh, “TOXGB: Teamwork Optimization Based XGBoost model for early identification of post-traumatic stress disorder,” Cognitive Neurodynamics **16**(4), 833–846 (2022).
6. Singh, S. Gupta, L. Goel, A.K. Agarwal, and S.K. Dargar, “Archimedes optimization-based Elman Recurrent Neural Network for detection of post-traumatic stress disorder,” Biomedical Signal Processing and Control **90**, 105806 (2024).
7. J. Hu, C. Zhao, C. Shi, Z. Zhao, and Z. Ren, “Speech-based recognition and estimating severity of PTSD using machine learning,” Journal of Affective Disorders **362**, 859–868 (2024).
8. J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Rizzo, and L. Morency, “The distress analysis interview corpus of human and computer interviews,” in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14) (European Language Resources Association, Reykjavík, Iceland, 2014), pp. 3123–3128.
9. Choubey, M.K. Pandey, and A.K. Dubey, “An Efficient Method to Recognize and Separate Patient’s Audio from Recorded Data,” in Proceedings of the Institute of Electrical and Electronics Engineers Conference (IEEE, Delhi, India, 2022), pp. 1–4.
10. W. Koeing, “A new frequency scale for acoustic measurements,” Bell Telephone Laboratories Record **27**, 299–301 (1949).
11. V. Tiwari, “MFCC and its applications in speaker recognition,” International Journal on Emerging Technologies **1**(1), 19–22 (2010).
12. D. Karaboga, and B. Akay, “A modified Artificial Bee Colony (ABC) algorithm for constrained optimization problems,” Applied Soft Computing **11**(3), 3021–3031 (2011).