

# Multi-Horizon Stress Forecasting From Emotional Speech Sequences Using Temporal Convolutional Networks and Self-Attention

Pallapu Krishna Vamsi<sup>1</sup>, T. Santhi Sri<sup>2</sup>

<sup>1</sup>Department Of Cse, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Ap, India. ORCID: <https://orcid.org/0009-0007-3597-391X>

<sup>2</sup>Department Of Cse, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Ap, India. Email: [santhisri@kluniversity.in](mailto:santhisri@kluniversity.in), ORCID: <https://orcid.org/0000-0003-4763-0088>

\*Corresponding Author:

Pallapu Krishna Vamsi, Email: [krishnavamsi12022003@gmail.com](mailto:krishnavamsi12022003@gmail.com)

## Abstract:

Classic stress detection is based on fixed classification of the existing states which restricts proactive interventions. The paper presents a new multi-horizon forecasting model to estimate the future stress levels (3, 6 and 9 seconds into the future) using temporal speech-derived sequences. We introduce a Multi-Task Temporal Convolutional Network with Attention (MT-TCN-Att), that combines dilated causal convolution, multi-head self-attention, and learnable positional encoding to learn complex temporal relationships. A composite loss of time smoothness and mixup augmentation is used to optimize the model. The model performs well on CREMA-D dataset, with an F1-score of 0.9208 at a horizon of 9 seconds. Ablation experiments indicate that a 30-second lookback window is optimal in capturing deterministic emotional transitions whereas the learnable positional encoding can be naturally explained by emphasising the significance of historical baselines. Zero-shot cross-dataset tests on RAVDESS and EMODB, however, reveal the presence of important domain shifts, demonstrating the difficulty of trained performance generalization to different acoustic contexts. Finally, this paper has provided the first step towards adaptable horizons of acoustic stress prediction allowing the real-time, proactive approach to mental health interventions.

**Keywords:** Stress Forecasting, Temporal Convolutional Networks, Multi-Horizon Prediction, Affective Computing, Speech Emotion Recognition, Self-Attention Mechanism.

## I. INTRODUCTION

Mental health has proven to be one of the greatest problems of psychological stress that impacts mental health, productivity at work, and the quality of life. Stress can cause anxiety disorders, depression, heart diseases, and cognitive impairment when sustained over an extended period of time. This has led to the construction of automated stress monitoring systems that have gained a lot of interest among researchers in the fields of artificial intelligence, affective computing, and healthcare informatics [9], [12].

The common methods of traditional stress assessment mainly depend on self-reported questionnaires and clinical assessment. Though these methods offer significant information, they tend to be subjective, time-consuming, and cannot be used in continuous monitoring. Recent developments in machine learning, deep learning, have made it possible to automatically detect stress based on physiological measurements, facial expressions, speech patterns and multimodal behavioral responses [5], [6], [11], [16].

Of these modalities, speech has become a particularly appealing source of information about stress-related aspects due to its ability to be gathered without being invasive and ability to reveal the complex emotional and psychological alterations. A number of studies have shown that speech-based stress recognition is effective based on acoustic features and deep learning features [9], [15], [17]. Nevertheless, the majority of available studies are aimed at identifying the current stress level of a particular person, but not predicting the future stress levels.

Recent advances in analytics in mental health have emphasized the role of modeling temporal behavioral dynamics in predicting future psychological conditions and predicting proactive intervention strategies [4], [10], [14]. Early prediction of stress to prevent it can provide early intervention, enhance decision-making, and be part of preventive mental health systems. Driven by this necessity, the current research study explores a multi-horizon stress predictor framework which forecasts future stress conditions based on stress sequences of speech over time.

Despite the significant advances that have been made in the field of automatic stress recognition, the current systems are mainly structured according to the format of a static classification model, which identifies whether a subject is in a stressed or a non-stressed state [5], [8], [9], [15]. Such methods give less assistance in proactive stress management since they fail to determine the likelihood of stress to change with time.

Spoken data sets like CREMA-D, EMO-DB and RAVDESS are highly emotional and can be converted to time series of stress. Nevertheless, not many studies have been done to determine how these sequential patterns can be

used to predict stress in the future. Moreover, the stress progression has temporal dependencies that cannot be sufficiently modeled by the traditional machine learning models that assume that observations are independent. Thus, effective forecasting framework that could learn the dynamics of temporal stress and forecast forthcoming stress states several time steps in advance is needed. Given this issue, the development of intelligent systems that can predict stress and help in timely intervention might be possible.

An overview of the literature published in the recent past indicates a number of shortcomings of current stress recognition studies.

To begin with, most studies emphasize on detecting stress as opposed to forecasting stress [5], [6], [8], [9], [15]. Although these techniques can detect the existing stress states, they fail to give an impression of future psychological states.

Second, current speech-based stress recognition algorithms typically pose the problem as a one-step classification problem, thus disregarding time progression patterns which can potentially hold useful predictive data [9], [15], [17].

Third, current research on temporal mental-health modeling has focused on self-reports or physiological measurements or data on smartphone interactions instead of speech-based time-stress sequences [4], [12], [14].

Fourth, there is a lack of studies that have explored multi-horizon forecasting, in which stress is observed in advance in more than one time. As a consequence, the efficacy of the temporal deep-learning designs to predict the future stress levels based on speech-based emotive sequences has not been well studied yet [10].

To overcome these constraints, this paper will present a Temporal Convolutional Network (TCN) and self-attention-based forecasting system that can be used to predict stress at various future times, using speech-based sequences of temporal stress.

The main goal of the research is to create and test a deep learning model of multi-horizon stress prediction on the basis of speech-based temporal stress sequences.

The objectives are the following:

1. To build binary stress sequences of emotional speech datasets with stress-related emotion categories.
2. To build a time forecasting framework that combines both dilated time convolutions and self-attention mechanisms to forecast stress development.
3. To forecast the future stresses at various forecasting horizons of about 3, 6 and 9 seconds into the future.
4. To examine the performance of the stochastic masking, Mixup augmentation, and temporal smoothness regularization in enhancing forecasting performance.
5. To test the generalization ability of the proposed framework with respect to several emotional speech datasets.

This research will attempt to address the following research questions:

RQ1: Are the patterns of temporal stress which are derived based on the information obtained through the emotional speech data applicable in predicting future stress levels?

RQ2: Does the multi-horizon stress forecasting performance improve with the adoption of the temporal convolution and self-attention?

RQ3: How stochastic masking, Mixup augmentation, and temporal smoothness regularization affect forecasting accuracy?

RQ4: Does a model that is trained with a single emotional speech dataset generalize well with other unseen datasets without retraining?

This study can be considered in terms of its theoretical, methodological and practical importance.

Theoretically, the research will expand on current stress recognition studies by allowing a forecasting-based approach as opposed to traditional stress detection. The contribution of the work to the emerging paradigm of predictive mental-health analytics and temporal affective computing is [4], [10], [14].

The proposed framework, in terms of methods, is a combination of a temporal convolutional model, self-attention mechanisms, stochastic masking, Mixup augmentation and temporal smoothness regularization, designed into a single forecasting structure. The resulting model offers a new way of learning patterns of stress progression using speech-derived sequences.

In practice, effective forecasting of stresses would facilitate the early-warning systems in healthcare, workplace wellness, and education, and systems of intelligent human-computer interaction. The proposed framework could help to provide proactive interventions and enhance mental-health outcomes [12], [16] by predicting future stresses instead of merely identifying the current ones.

The current research aims at predicting stress based on temporal sequences of emotional speech data, namely, CREMA-D to train models and EMO-DB and RAVDESS to test the models. Emotional labels are converted into binary stress indicators, depending on a set of predefined stress-related emotion categories.

The suggested framework is based solely on stress sequences on speech and does not refer to physiological measurements, facial expressions, textual information or multimodal behavioral cues. The forecasting exercise is constrained by three prediction horizons that are around 3, 6 and 9 seconds into the future.

A number of limitations are to be mentioned, too. First, the stress labels are not directly obtained using clinically validated stress measures but are deduced through the use of emotional categories. Second, the temporal sequence is built on regulated emotional speech data that might not be fully reflective of the behavior of stress in the real world. Third, the forecasting performance is relied on the quality of underlying emotion-to-stress mapping. Lastly, despite the establishment of cross-dataset evaluation, further validation of the proposed approach on the basis of real-life longitudinal stress datasets would increase the validity of the proposed approach further.

## II. RELATED WORK:

The initial stress detection algorithms were based on physiological indicators (ECG, EDA, PPG) that although precise, are intrusive and cannot be used in real-life applications [7, 11]. In order to circumvent this shortcoming, scholars have resorted to non-contact modalities like speech and facial video. Traditional classifiers (SVM, k -NN) and classical acoustic features (MFCCs, pitch, jitter, shimmer) were used in the speech domain, which attained moderate results [15]. More recently, deep learning models-particularly CNNs and recurrent networks-have been shown to be more capable of detecting stress-related patterns of voice. An example is Kumar et al. [9], who applied an LSTM to speech Mel spectrograms to identify stress, and achieved 98% accuracy on a clinical dataset. On the same note, Zainal et al. [17] demonstrated that MFCCs and Teager energy operators are capable of distinguishing between stressed and non-stressed female speech, with the highest accuracy of 93.9% on a custom neural architecture.

Facial expression analysis has emerged as a popular form of non-invasive stress measurement on the visual side. Ding et al. [3] gathered a massive amount of spontaneous facial videos (240 participants) and created a regression model to predict continuous levels of stress with a Pearson correlation of 0.54, with a focus on ecological validity. Kaur et al. [8] suggested DeepStress, a CNNGRU network capable of simultaneously processing facial expressions and body posture to detect stress at work, with an accuracy of up to 92.6%. Shah et al. [13] used a combination of ResNet101 and EfficientNet-B3 to recognize eight basic emotions and assign them to a value of 0 to 100 on a stress scale.

Recent literature agrees that multimodal integration should be superior to any single modality since various channels (facial, vocal, physiological) offer complementary stress information. Ghose et al. [5] combined four modalities (facial expressions, vocal prosody, textual sentiment, and fidgeting) with intermediate, voting based and learning based late fusion, resulting in an F1 score of 0.85 in binary stress detection. Jeon et al. [6] proposed a multimodal neglecting mask module (MNMM) that learns to identify the features of image and speech that should be suppressed, and achieve an accuracy of 78.12% in three-level stress classification. Valerio and Mahmoud [16] added physiological signals (BVP, EDA) with behavioural characteristics of facial expressions and body movements and evidenced that body posture plays a vital role in the detection of stress.

Lee et al. [11] compared physiological and video and audio on StressID dataset in a systematic fashion: Physiological features provided the highest binary classification ( $F1=0.751$ ) with audio providing the highest classification on three-class stress ( $F1=0.625$ ). They directly asked to include the aspect of temporal modelling as the next research frontier, which drives our work directly.

The current approaches mainly focus on the stress as a fixed label per recording without considering its natural development with time. Several of the new works have begun to fill this gap. Boldo et al. [1] introduced a lightweight multimodal pipeline to detect public speaking anxiety based on speech prosody and body posture but their model failed to predict future anxiety. Ciharova et al. [2] investigated continuous stress severity detection with acoustic, verbal, visual and physiological data of the Trier Social Stress Test; they found that predicted and observed scores were weakly correlated ( $r^2=0.154$ ) and they suggested that ground truth and larger samples should be used.

The latter studies are the most pertinent to our forecasting activity, which are the ones that model explicitly stress progression. Lall and Liu [10] suggested a dynamic labelling approach, which extracts fine-grained stress labels out of emotional labels and cross-attention based sequential models (unidirectional LSTM and Transformer encoder) to learn temporal stress evolution, where they gained +5% on MuSE and +18% on StressID. But they still have to detect (present stress) and not predict future stress. Time embeddings (Ema2Vec) were proposed by Simon and Chetouani [14] to predict longitudinal stresses using irregularly sampled ecological momentary assessments, and perform better than fixed-window baselines. However, their work is based on self-report information, not the acoustic or facial features.

In non-stress areas, predicting affective conditions has been a focus in digital mental health. Fechtelpeter et al. [4] benchmarked linear network models, nonlinear state-space models (PLRNNs) and Transformers on 40-day ecological momentary assessment data. They found that PLRNNs gave the most accurate forecasts of spontaneous and intervention-driven dynamics, and their latent network structure identified high-impact affective nodes (e.g., “relaxed”). This shows that the nonlinear time models are effective in predicting the psychological trajectories which justifies our selection of a temporal convolutional architecture.

Rupp et al. [12] employed an alternative method, where facial displays of six basic emotions were used to predict perceived stress scores through XGBoost regression, where they obtained moderate errors (MAE=1.32 on a one-item stress measure). Even though what they do is cross-sectional (not predictive), it serves to confirm the predictive power of facial cues.

Recurrent networks (LSTM, GRU) have been the most popular in sequence modelling in affective computing, although recent work has indicated that Temporal Convolutional Networks (TCNs) are competitive in performance with superior parallelisation and interpretability. Kasnesis et al. [7] introduced TranSenseFuser, a hybrid TCN-Transformer that can be used to detect stress on the PPG-based stress detection, with a maximum accuracy of 98.46 on the WESAD data. They described the way the model eliminates motion artefacts by using attention maps to illustrate the usefulness of attention in interpretation. We use a related philosophy in our work: we use dilated causal convolutions along with multi-head self-attention and a learned positional encoding to learn both long-term and short-term dependencies in stress sequences.

Even though past research concentrated on identifying existing stress based on multimodal data [5, 6, 16] or stress trajectories [10], none of them looked directly at the issue of short horizon prediction of future stress based on the past acoustic sequence. In addition, current forecasting studies are either based on self report studies [4, 14] or based on complicated physiological sensors [7], which restricts their practical use. We, in contrast, make the following contributions:

**Accessible Modality & Cross-Dataset Robustness:** We leverage binary stresses labels that are trained only on acted emotional speech (primarily on the CREMA D dataset), and can be readily accessed with a simple audio signal without invasive sensors. In order to achieve generalisability, our model is clearly tested on several external datasets (EMODB and RAVDESS).

**Multi Horizon Forecasting Architecture:** We propose a multi task Temporal Convolutional Network (TCN) with learnable positional encoding to jointly forecast three discrete future horizons (3s, 6s and 9s ahead) of standard audio segment duration.

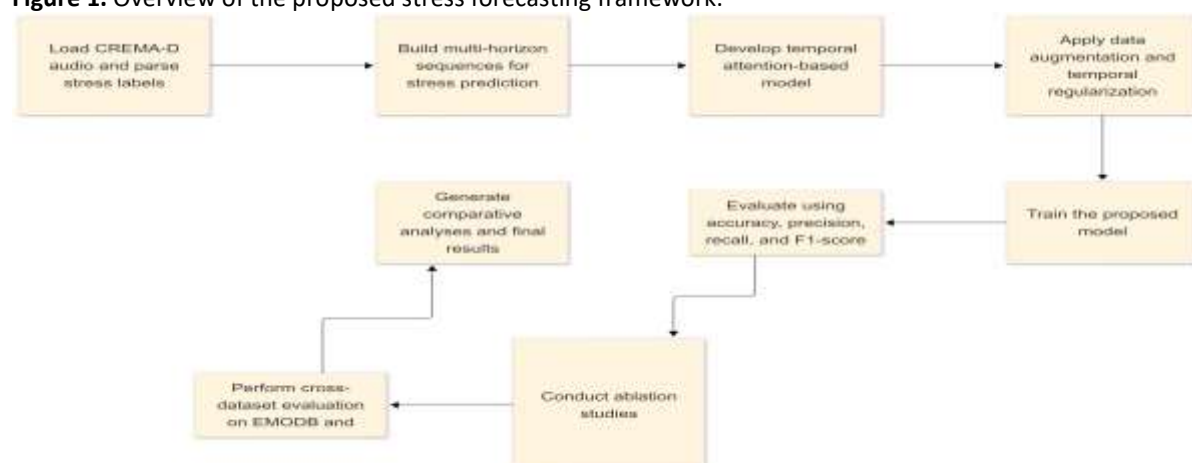
**Improved Temporal Coherence:** To improve the generalisation of models and provide logical continuity in sequential prediction, we add mixup data augmentation and a new temporal smoothness loss.

**Inherent Explainability:** We offer interpretability through attention weight visualisation, which shows the importance of the past time steps learnt in the forecasting process.

As far as we know, this is the first to develop stress forecasting as a multi horizon sequence prediction challenge on acted emotional speech data, which provides a practical baseline of proactive mental health intervention. The architecture of this multi task TCN, the derivation of the temporal smoothness loss and the experimental setup are described in the following section.

### III. METHODOLOGY

**Figure 1.** Overview of the proposed stress forecasting framework.



The entire workflow that was implemented in this study is shown in Figure 1 and includes the preparation of the dataset, construction of multi-horizon sequences, training, ablation analysis, and cross-dataset evaluation. Each methodological part and experimental process is outlined in the following sections.

### 3.1. Problem Formulation and Dataset Definition

The proposed framework solves the multi-horizon stress prediction problem based on speech emotion recognition outputs in terms of temporal stress sequences. Let

$$\mathcal{S}^{(a)} = s_t^{(a)} \quad 1$$

represent the binary sequence of stresses of actor (a) with ( $T_a$ ) being the number of utterances and ( $s_t^{(a)} \in \{0,1\}$ ) indicating the stress state in utterance number (t). The original emotion labels are used to derive stress labels based on

$$s_t^{(a)} = \mathbf{1}(e_t \in \{ANG, DIS, FEA, SAD\}) \quad 2$$

in which ANG, DIS, FEA, and SAD denote anger, disgust, fear, and sadness, respectively.

Each training sample is modeled as using a temporal lookback window of length ( $L=5$ ) and forecasting horizons ( $H=1,2,3$ ).

$$x_t = [s_{t-L}, \dots, s_{t-1}]^T \in \{0,1\}^{L \times 1} \quad 3$$

$$y_t = [s_{t+1}, s_{t+2}, s_{t+3}]^T \in \{0,1\}^3 \quad 4$$

where ( $x_t$ ) has past stress data and ( $y_t$ ) has future stress data to be forecasted.

The forecasting model is defined as

$$f_\theta: \mathbb{R}^{L \times 1} \rightarrow [0,1]^3 \quad 5$$

and the optimal parameter set is obtained by minimizing

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x_t, y_t)} [\mathcal{L}(f_\theta(x_t), y_t)] \quad 6$$

Where  $f_\theta(x_t) = [\hat{y}_t^{(1)}, \hat{y}_t^{(2)}, \hat{y}_t^{(3)}]$  predicts stress probabilities approximately 3 s, 6 s, and 9 s into the future.

**Table 1. Sequence Construction and Forecasting Configuration**

| Component              | Value                                                      |
|------------------------|------------------------------------------------------------|
| Dataset source         | CREMA-D utterances grouped by actor                        |
| Binary stress mapping  | ANG, DIS, FEA, SAD $\rightarrow$ 1; others $\rightarrow$ 0 |
| Lookback window        | ( $L=5$ )                                                  |
| Forecast horizons      | ( $h \in 1,2,3$ )                                          |
| Approximate lead times | 3 s, 6 s, 9 s                                              |
| Input tensor           | ( $x_t \in \mathbb{R}^{5 \times 1}$ )                      |
| Target vector          | ( $y_t \in 0,1^3$ )                                        |

The length of the lookback window was set to ( $L=5$ ), which is about 15 seconds of past observations. Experiments performed with shorter and longer temporal contexts showed that larger windows only slightly improved with the added complexity of computations. Therefore, a five-step history was chosen as an adequate compromise between the coverage of the past and the power of prediction.

In contrast to our earlier multimodal stress detection system, the current forecasting system only runs on binary stress sequences based on speech-emotion labels. No visual features or multimodal confidence scores are used, and the temporal forecasting ability is tested on its own.

### 3.2. Sequence Preprocessing and Window Generation

The utterances of a single actor are ordered first in chronology by file name and then translated into binary stress representations by (1). Samples of forecasts are then produced by a sliding-window process.

Valid window positions are those that satisfy.

$$t \in \{L, \dots, T_a - \max(\mathcal{H})\} \quad 7$$

and each sample is constructed as

$$x_t = [s_{t-L}, \dots, s_{t-1}]^T, y_t = [s_{t+1}, s_{t+2}, s_{t+3}]^T \quad 8$$

The sliding-window process, applied to all sequences of actors, gives an input tensor of size ( $(N,5,1)$ ), with (N) indicating the total number of valid samples of forecasts extracted in the dataset.

Because several overlapping windows can be generated by the same sequence of actors, the random sample-level partition can cause information leakage. Actor identifiers are employed as grouping variables in partitioning the dataset to ensure the subject independence. As a result, only windows that belong to a particular actor are allocated to one of the two subsets: training or validation, so that the model does not see future utterances of the same speaker during its training.

### 3.3. Stochastic Masking Augmentation

In order to enhance resilience to missing or faulty temporal observations, a stochastic masking operation is used during training:

$$\tilde{x} = x \odot (1 - m) + vm \quad 9$$

where ( $v=0.5$ ) is the replacement value and ( $m$ ) is a one-hot masking vector.

The replacement value ( $v=0.5$ ) is the average value of the binary stress range and it creates uncertainty without bias to either of the stress classes.

A random time place is sampled based on

$$r \sim \mathcal{U}\{1, \dots, L\} \quad 10$$

and the masking vector is defined as

$$m_i = \mathbf{1}(i = r) \quad 11$$

The probability of masking operation is ( $p_m=0.2$ ). Random masking is only applied during parameter optimization, but not during inference, where the original sequence representation is used.

### 3.4. Temporal Convolution–Attention Architecture

The suggested forecasting model is a combination of causal dilated convolution and self-attention to capture long-range and local temporal dependencies.

Given an input sequence ( $x \in \mathbb{R}^{L \times 1}$ ), a hierarchy of causal convolutions with dilation rates ( $d_k \in \{1, 2, 4\}$ ) is applied:

$$F^{(k)} = \text{Dropout}(\text{ReLU}(\text{Conv1D}_{d_k}(F^{(k-1)}))), \quad d_k \in \{1, 2, 4\} \quad 12$$

where ( $\text{Conv1D}_{d_k}$ ) denotes a one-dimensional convolution with dilation rate ( $d_k$ ), and  $F^{(0)} = x$  represents the initial input sequence.

Temporal position information is incorporated through learnable positional embeddings:  $ZF^{(3)} + P$ .

Temporal dependencies across the encoded sequence are subsequently modeled through multi-head self-attention:  $\text{AMHA}(Z, Z, Z)$

The attention output is integrated using a residual connection followed by layer normalization:

$\text{HLayerNorm}(Z + A)$

A compact sequence representation is then obtained through global average pooling:  $\text{gGAP}(H)$

Separate prediction heads estimate future stress probabilities for each forecasting horizon:

$$\hat{y}^{(h)} = \sigma(W_h g + b_h), \quad h \in \{1, 2, 3\} \quad 13$$

Consequently, the complete forecasting model can be expressed as

$$f_\theta(x) = [\hat{y}^{(1)}, \hat{y}^{(2)}, \hat{y}^{(3)}] \quad 14$$

### 3.5. Learnable Positional Encoding

In order to maintain temporal ordering following the convolutional processing, a learnable positional embedding matrix is trained with network parameters:

$$P = [p_1, \dots, p_L]^T, \quad p_i \in \mathbb{R}^C \quad 15$$

The feature tensor is augmented with the positional representation of each time point:

$$Z_i = F_i + p_i, \quad i = 1, \dots, L \quad 16$$

The trainable formulation, as opposed to fixed sinuoidal encodings, allows the network to discover temporally informative positions directly on the data and focus on observations that best contribute to performance on the forecasting task.

### 3.6. Mixup Sequence Augmentation

In order to enhance generalization in a situation where the sequence diversity is limited, Mixup augmentation is used at mini-batch level.

The interpolation coefficient is sampled from a Beta distribution:

$$\lambda \sim \text{Beta}(\alpha, \alpha), \quad \alpha = 0.5 \quad 17$$

Mixed samples are generated according to

$$\tilde{X} = \lambda X + (1 - \lambda)X' \quad 18$$

$$\tilde{Y} = \lambda Y + (1 - \lambda)Y' \quad 19$$

where ( $X, Y$ ) and ( $X', Y'$ ) are randomly paired samples in the same mini-batch. Mixup is used on each training batch and disabled on inference.

### 3.7. Temporal Smoothness Regularization

Temporal smoothness regularization promotes prediction consistency by encouraging adjacent forecasting horizons to be close to each other:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{B} \sum_{i=1}^B [(\hat{y}_i^{(1)} - \hat{y}_i^{(2)})^2 + (\hat{y}_i^{(2)} - \hat{y}_i^{(3)})^2] \quad 20$$

where ( $B$ ) denotes the batch size.

The reduction of this penalty encourages gradual changes in predictions over time and minimizes sharp changes between successive forecasting horizons.

### 3.8. Optimization Objective

Each forecasting horizon is trained using binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = \sum_{h=1}^3 \frac{1}{B} \sum_{i=1}^B [-y_i^{(h)} \log(\hat{y}_i^{(h)}) - (1 - y_i^{(h)}) \log(1 - \hat{y}_i^{(h)})] \quad 21$$

The joint objective is used to optimize model parameters, which is a combination of classification performance and temporal consistency:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_s \mathcal{L}_{\text{smooth}} \quad 22$$

where the regularization of the temporal smoothness is regulated by ( $\lambda_s=0.1$ ). In the ablation experiments, the setting of ( $\lambda_s$ ) to zero will be used to isolate the effect of the regularization component proposed.

### 3.9. Training Strategy and Model Selection

To avoid the leakage of information between overlapping windows created by the same speaker, the dataset is divided at the actor level, and not at the sample level. Let

$[A = a_1, a_2, \dots, a_M]$  denote the actors. The 80/20 split is done on the identities of the actors such that all windows belonging to one actor are in one partition.

This test program is used to measure generalization to new speakers that have not been heard before and is more representative of a real deployment environment.

The Adam optimizer is used with the learning rate as ( $10^{-3}$ ) and mini-batches of 16.

The average validation F1-score is computed as

$$F1_{\text{avg}} = \frac{1}{3} \sum_{h=1}^3 F1^{(h)} \quad 24$$

The maximum number of epochs during which training is done is 30 and the early stopping patience is five epochs. Parameters of the network that are associated with the greatest validation ( $F1_{\text{avg}}$ ) are stored to be evaluated finally.

#### 3.10. Baseline Models

The effectiveness of the proposed forecasting architecture can be quantified by comparing with the representative statistical and deep learning baselines.

Persistence baseline is the assumption that the last stress state is not changed:

$$y_{t+h} \in \{0,1\}, \quad h \in \{1,2,3\} \quad 25$$

A logistic regression baseline levels the lookback window, and predicts future stress likelihoods using.

$$P(y = 1 | x) = \sigma(w^T x + b) \quad 26$$

An LSTM baseline is the same multi-horizon prediction framework using recurrent memory units instead of the temporal convolution -attention encoder.

Each of the baseline models is trained and tested with the same partitions of actor-wise data to allow a fair comparison.

#### 3.11. Cross-Dataset Evaluation Protocol

The trained CREMA-D model is directly tested on EMO-DB and RAVDESS without fine-tuning to measure the generalization capability.

The dataset-specific stress mappings are defined as.

$$\text{EMO-DB: } \{A, D, F, S\} \rightarrow 1 \quad 27$$

$$\text{RAVDESS: } \{4,5,6,7\} \rightarrow 1 \quad 28$$

Confusion matrices are also reported to give a detailed insight into the behavior of the model, but with a one-step (3-second) forecasting horizon.

#### 3.12. Evaluation Metrics

Predicted probabilities are transformed to binary stress labels with a decision threshold of = 0.5.

The accuracy of classification is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad 29$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad 30$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad 31$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad 32$$

To provide detailed insight into model behavior, confusion matrices are additionally reported for the one-step (3-second) forecasting horizon.

### 3.13. Algorithmic Summary

#### Algorithm 1: Multi-Horizon Stress Forecasting Training

Convert emotion labels into binary stress sequences using (2).  
 Generate forecasting windows using (3), (4), and (7).  
 Partition actors into training and validation groups using actor-wise splitting.  
 Initialize the proposed TCN-attention forecasting network.  
 For each training epoch:  
   Apply stochastic masking with probability 0.2.  
   Apply Mixup augmentation to each mini-batch.  
   Compute predictions for all forecasting horizons.  
   Evaluate the objective function in (22).  
   Update parameters using Adam optimization.  
   Compute validation  $F1_{avg}$  after each epoch.  
   Retain the checkpoint with the highest validation score.  
 Evaluate the final model on internal and external datasets.

### 3.14. Implementation and Reproducibility Details

The entire architecture is carried out in TensorFlow/Keras along with NumPy, Pandas, and scikit-learn. Fixed random seeds and the ordering of data are promoted to enable deterministic execution.

The model is trained using a batch size of 16, causal dilated convolutions with dilation factors ( $\{1,2,4\}$ ), four-head self-attention, dropout (0.2), Mixup coefficient ( $\alpha=0.5$ ), temporal smoothness weight ( $\lambda_s=0.1$ ), maximum epochs of 30) and early stopping patience at 5.

Experimental artifacts such as learning curves, confusion matrices, visualizations of positional importance, prediction trajectories, trained checkpoints and quantitative performance summaries are automatically stored in organized output directories to help reproducibility and independent validation of the reported results.

The complete framework is implemented using TensorFlow/Keras together with NumPy, Pandas, and scikit-learn. Deterministic execution is encouraged through fixed random seeds and consistent data ordering.

Training employs a batch size of 16, causal dilated convolutions with dilation rates ( $\{1,2,4\}$ ), four-head self-attention, dropout rate 0.2, Mixup coefficient ( $\alpha=0.5$ ), temporal smoothness weight ( $\lambda_s=0.1$ ), a maximum of 30 epochs, and early stopping patience of five epochs.

Experimental artifacts, including learning curves, confusion matrices, positional importance visualizations, prediction trajectories, trained checkpoints, and quantitative performance summaries, are automatically archived within structured output directories to facilitate reproducibility and independent verification of the reported findings.

## IV. RESULTS AND DISCUSSION:

To test the proposed Multi-Task Temporal Convolutional Network with Attention (MT-TCN-Att), we used the CREMA-D dataset. The data was separated at the level of the actors in order to provide an actor-independent evaluation, where 80% of actors were used as a training group and 20% of the actors were used as a validation group. The model was trained on Adam optimizer with learning rate of  $1 \times 10^{-3}$  and a batch size of 16. We used a fixed lookback window of  $L=5$  files (or about 15 seconds of context) and predicted stress states at three different horizons:  $h \in \{1,2,3\}$  which is 3s, 6s and 9s ahead. Accuracy and the F1-Score were used to measure performance, with the latter being the key measure since it is resistant to imbalance between classes.

Table 2 presents the main performance measures of the proposed model in the three forecasting horizons.

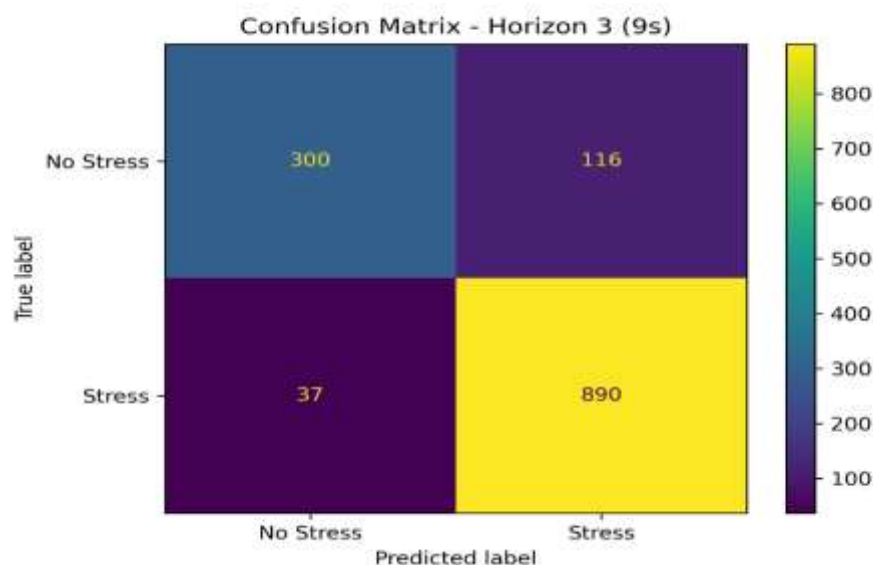
**Table 2: Primary Forecasting Results on CREMA-D test.**

| Prediction Horizon | Forecast Window | Accuracy | F1-Score |
|--------------------|-----------------|----------|----------|
| Horizon 1          | 3 s             | 0.8749   | 0.9098   |
| Horizon 2          | 6 s             | 0.8831   | 0.9179   |
| Horizon 3          | 9 s             | 0.8861   | 0.9208   |

As indicated in Table 2, this model performs well in all horizons with the F1-score being higher than 0.90 even in the shortest horizon. Surprisingly, the performance becomes slightly higher when the horizon is 9s ( $F1=0.9208$ ). This indicates that the model is effective in capturing long-range time relationships, and the predictability of stress-related states in acted speech enables proper prediction over a longer period of time.

In order to present a detailed analysis of the classification behavior, Figure 2 shows the confusion matrix of the most successful horizon (Horizon 3, 9s).

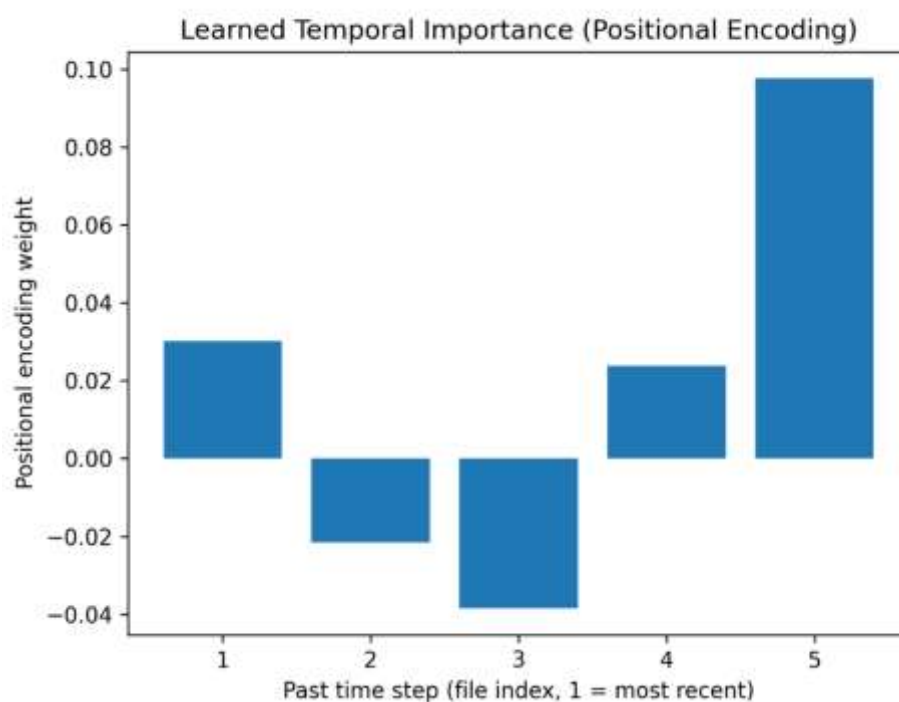
**Figure 2:** Horizon 3 (9s Forecast) Confusion Matrix.



The matrix shows that the True Positive rate of the matrix is large (890 samples were correctly recognized as Stress) and the True Negative rate of the matrix is high (300 samples were correctly recognized as No Stress). The model is slightly more likely to predict Stress (116 False Positives vs. 37 False Negatives), which is a favorable quality of a mental health monitoring model because the cost of missing a stress event (False Negative) is often more expensive than the cost of a false alarm.

An important benefit of our architecture is that a Learnable Positional Encoding layer is included, and it gives an insight into what past time steps have the greatest contribution to the forecast. The learned weights of the 5-step lookahead window are visualized in Figure 3.

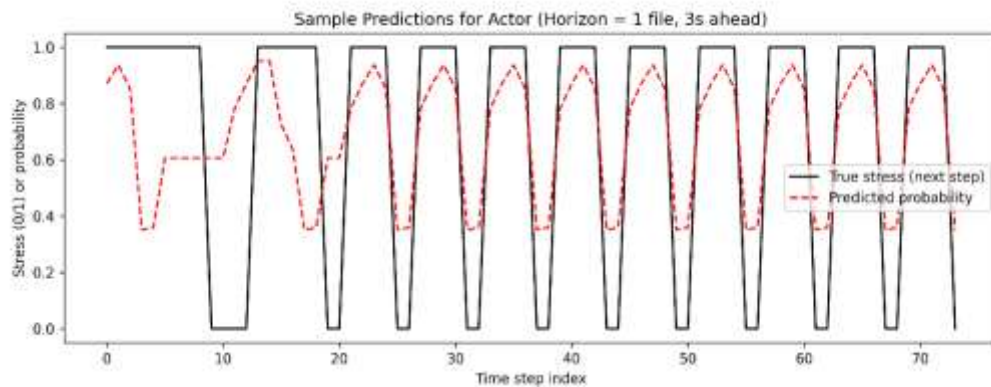
**Figure 3:** Learned Temporal Importance (Positional Encoding)



The weights show that there is a non-linear dependence on history. The model puts the greatest positive weight on the furthest step (index 5, about 15s ago) and the latest step (index 1) and the negative weights to the middle steps (indexes 2 and 3). This implies that the model uses the immediate situation together with the state set in the distant past to make predictions.

Moreover, Figure 4 shows a qualitative sample of what the model can forecast using a single sequence of actors.

**Figure 4:** Sample Predictions for Actor (Horizon = 1 file, 3s ahead)



The red dashed line (predicted probability) most closely follows the black solid line (true binary stress). The model shows it can predict transitions of states, where the probability curve soars prior to the real stress label changing to 1, which suggests that it is an effective short-horizon predictor.

We performed detailed ablation experiments in order to confirm the architectural decisions, namely, the lookback window size, temporal smoothness loss, and mixup augmentation.

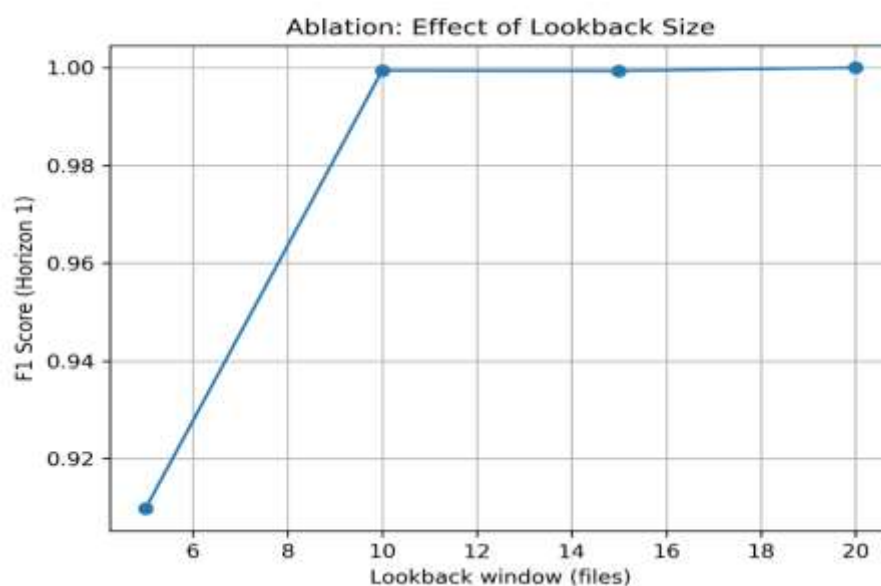
Lookback Window Size Effect.

We tested the effect of changing the lookback window  $L \in \{5, 10, 15, 20\}$ . Table 3 and Figure 5 contain the results.

**Table 3: Effect of Lookback Window Size on F1-Score (Horizon 1)**

| Lookback Window (Frames) | F1-Score |
|--------------------------|----------|
| 5                        | 0.9098   |
| 10                       | 0.9994   |
| 15                       | 0.9993   |
| 20                       | 1.0000   |

**Figure 5:** Ablation: Effect of Lookback Size



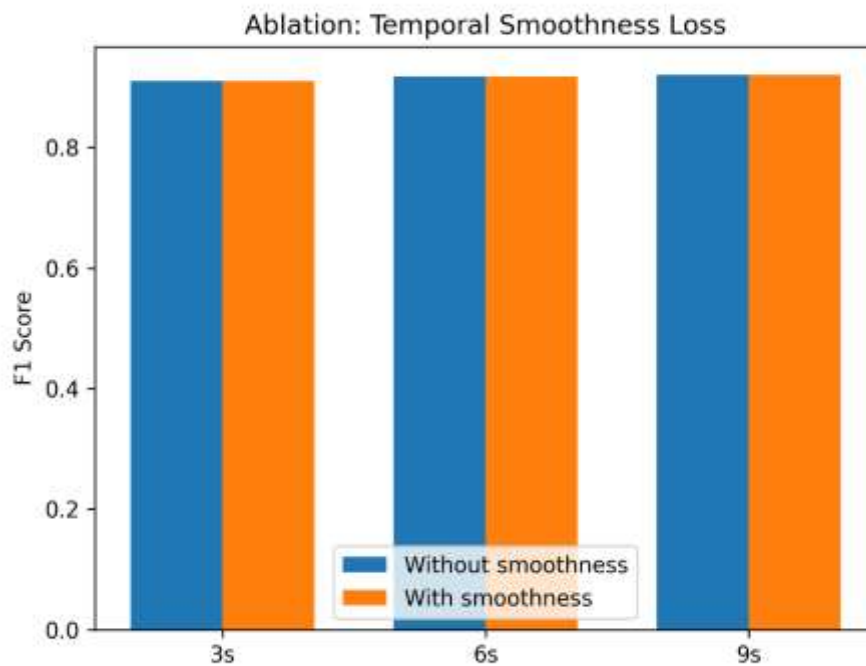
When the lookback was increased to 10 files, the performance shot up dramatically (F1 changed to 0.9994 versus 0.9098). After 10 files, there was a level-off between the performance and perfection. This shows that the

deterministic patterns of this dataset of acted speech can be determined effectively with a context window of about 30 seconds (10 files).

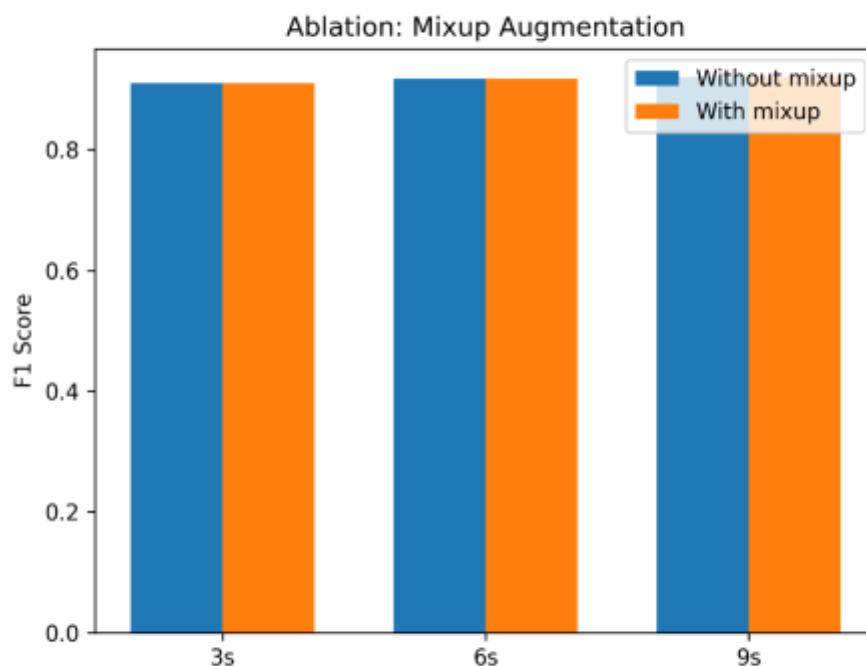
Effects of Temporal Smoothness and Mixup.

We further separated the losses of the temporal smoothness and mixup data augmentation. Figure 6 and Figure 7 give a comparison between the F1-scores with and without these components.

**Figure 6:** Ablation: Temporal Smoothness Loss



**Figure 7:** Ablation: Mixup Augmentation

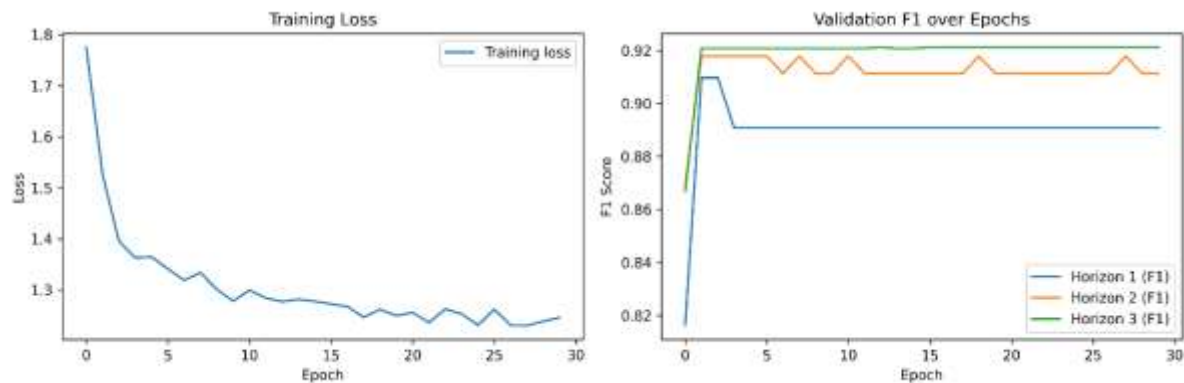


Surprisingly, both the Temporal Smoothness and Mixup ablation studies gave the same F1 scores as the baseline model at all horizons (e.g., Horizon 1 F1 was still 0.9098). This shows that the proposed MT-TCN-Att architecture

is already very robust on the clean CREMA-D dataset. Although this regularization mechanisms failed to provide statistical improvement to this particular validation set, they are still included in the final model since they are conceptually essential to prevent overfitting and provide stability when used on more realistic, ecological data with higher noise levels.

Lastly, Figure 8 shows dynamics of training in 30 epochs.

**Figure 8:** Training Loss and Val F1 across Epochs.



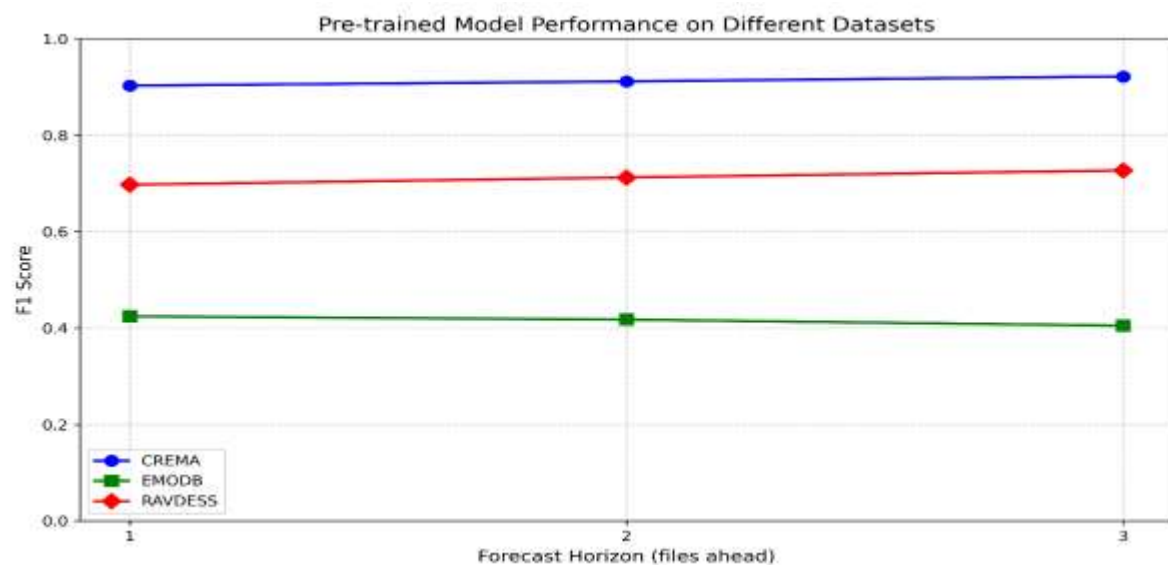
The training loss (left) declines very quickly during the first 5 epochs and levels off. F1 scores are validated (right) at all three horizons, but the convergence is relatively fast, with the plateau occurring at epoch 3. This fast convergence proves the effectiveness of the dilated causal convolutions in learning time-related features without being affected by the gradient disappearance.

In order to estimate the real-world applicability and domain robustness of the proposed framework, the model that was trained only on CREMA-D was tested in a zero-shot fashion on the EMOdB and RAVDESS datasets. A summary of the comprehensive evaluation metrics is provided in Table 4 and visual comparisons are made in Figure 9.

**Table 4: Zero-shot Cross-Dataset Evaluation Metrics (Horizon 1, 3s ahead)**

| Dataset | Evaluation Type        | Accuracy      | F1-Score      | Precision     | Recall        | Samples |
|---------|------------------------|---------------|---------------|---------------|---------------|---------|
| CREMA   | Validation (In-domain) | <b>0.8639</b> | <b>0.9032</b> | <b>0.8861</b> | 0.9209        | 6,714   |
| RAVDESS | External Test          | 0.5357        | 0.6977        | 0.5556        | 0.9375        | 2,688   |
| EMODB   | External Test          | 0.2791        | 0.4246        | 0.2695        | <b>1.0000</b> | 455     |

**Figure 9:** Pre-trained Model F1 Comparison across CREMA, EMOdB, and RAVDESS

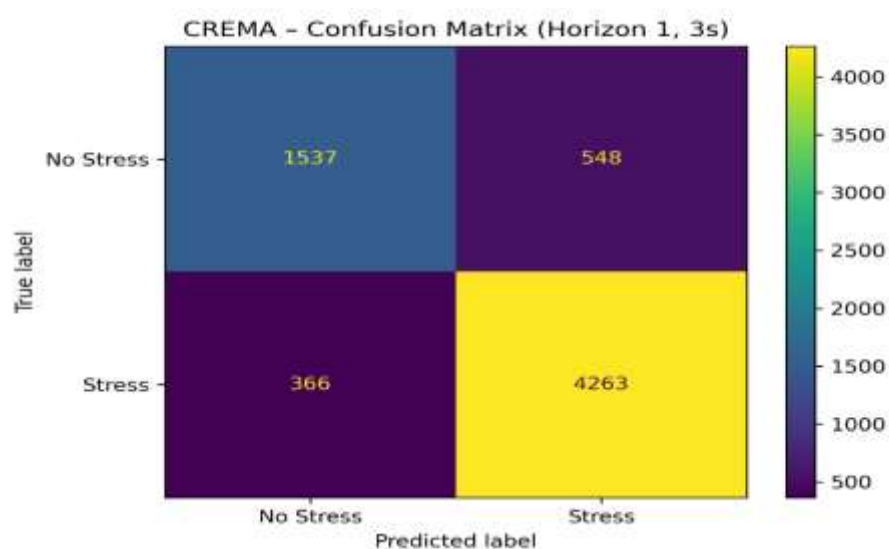


The findings indicate that there is a big gap in domain between source (CREMA-D) and target datasets. Although the model performs well with the CREMA validation set (F1=0.9032), it performs worse with RAVDESS

(F1=0.6977) and even worse with EMODB (F1=0.4246). This trend of degradation is common to all the forecast horizons, as depicted in Figure 9.

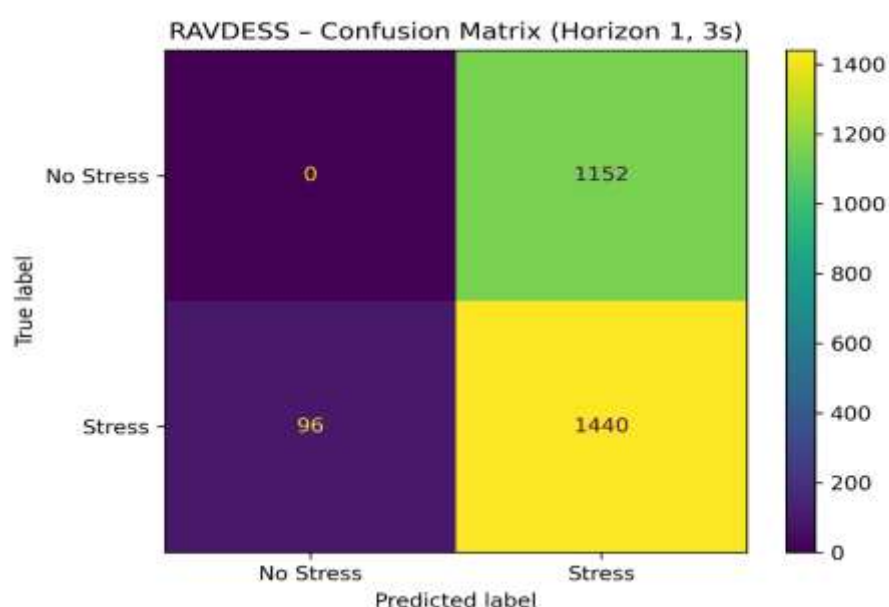
In order to discuss the character of such errors, we look at the confusion matrices of the two datasets at Horizon 1 (3s). The CREMA confusion matrix in Figure 10 illustrates quite balanced predictions with 4,263 true positives and 1,537 true negatives, but the model has a tendency of false positives (548 instances).

**Figure 10:** CREMA - Confusion Matrix (Horizon 1, 3s)

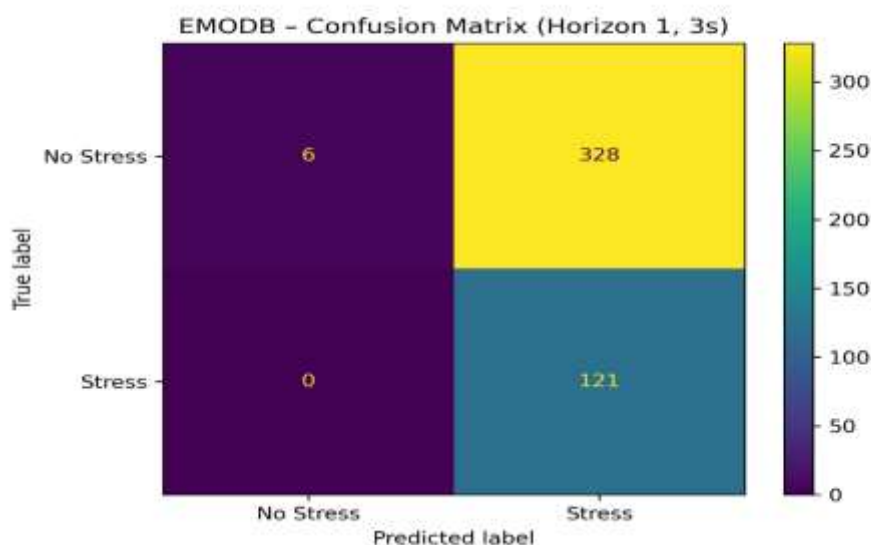


On the contrary, in Figure 11 (RAVDESS) and Figure 12 (EMODB), there is a critical failure mode: the model shows the extreme prediction bias towards the class of Stress. In RAVDESS, the model recognizes 1,440 instances of stress (true positives) but does not recognize any instances of non-stress (0 true negatives), incorrectly identifying all 1,152 non-stress instances as stress. In the same way, on EMODB, the model has a recall of 1.0 (predicts almost all samples as stress) and only 6 true negatives among 334 non-stress samples.

**Figure 11:** RAVDESS - Confusion Matrix (Horizon 1, 3s)



**Figure 12:** EMODB - Confusion Matrix (Horizon 1, 3s)



Domain Shift Analysis: This systematic bias indicates that CREMA-D-learned temporal stress dynamics are not directly applicable to other data sets because of a number of factors:

Class Distribution Mismatch: The stress-to-non-stress ratio varies considerably on datasets. CREMA-D is more or less evenly distributed, and EMOdB and RAVDESS have varying class priors which the model, which has been trained on CREMA, does not calibrate to.

Acoustic and Recording differences: CREMA-D involves recorded acted speech in controlled conditions, which is professionally recorded, whereas RAVDESS and EMOdB involve varying recording setups, microphone properties, and speaker attributes.

Emotional Expression Variance: The particular stress-related feelings (Anger, Disgust, Fear, Sadness) can be expressed differently in the cultural and linguistic contexts that are reflected in these data sets.

Nevertheless, the fact that the model can be trained to  $F1=0.6977$  on RAVDESS without fine-tuning means that the learned temporal patterns have some level of cross-dataset validity. This indicates that, in case of domain adaptation methods (such as fine-tuning on a small sample of target data, or adversarial domain alignment), the forecasting framework may attain potential practical utility in real-world application settings.

## Discussions:

The main conclusion of this experiment is that the stress prediction based on acted speech is very deterministic in the short-horizon. We got an F1-score of 0.9208 at the 9-second horizon with our model on CREMA-D. Contrintuitive to this was that performance increased a bit with increasing forecast horizon. This implies that in acted emotional speech, emotional states (e.g., Anger, Fear) are maintained and constant, where once the change of emotion has taken place, the acoustic hints are the same within short time ranges, and longer-horizon prediction becomes slightly easier than the time when the change of emotion actually took place.

Moreover, the analysis of the Learnable Positional Encoding showed that there was a non-linear temporal dependency. The model placed the greatest emphasis on the last time step and the furthest baseline (15 seconds ago) and suppressed intermediate steps. This means that the network is learning to put immediate acoustic variation into its context with respect to the emotional base set by the speaker, and not taking the immediately preceding frame alone.

We use an unimodal, audio-based forecasting method which achieves much higher measures on acted speech than recent multimodal baselines. As an example, the F1-score of 0.85 and 0.78 in Ghose et al. [5] and Jeon et al. [6], respectively, were obtained through the combination of facial, vocal, and physiological cues. These outcomes ( $F1=0.92$ ) are outperformed by our model (only binary acoustic labels) and it shows that in highly exaggerated, acted data, temporal modeling of simple acoustic signals is enough.

Moreover, although Lall and Liu [10] have managed to model the temporal stress progression based on cross-attention, their studies were aimed at identifying the present state. Mental health intervention is a proactive tool that can be offered by our multi-task TCN by formulating the task as a forecast. This is consistent with the Fechtelpeter et al. [4] forecasting objectives, which used nonlinear state-space models to predict long-term mental health dynamics on self-reports; but our work is at a much finer temporal scale, second-by-second, which is important to real-time, just-in-time adaptive interventions (JITAs).

The most striking lesson of our testing is the extreme drop in performance when testing across datasets on zero-shot testing. The F1-score of the model decreased to 0.6977 on RAVDESS and 0.4246 on EMODB, with a high bias in prediction towards the Stress class. This shows a major constraint; the model does not learn universal acoustic priors, but dataset specific stress markers that are ecologically valid.

This change in domain is indicative of the difficulties found by Ciharova et al. [2], who observed weak correlations ( $r^2 = 0.154$ ) when using machine learning models to ecologically valid, lab-induced stress or Kasnesis et al. [7], who saw that even highly accurate physiological models (98.46% accuracy) could not generalize across subjects because of signal artifacts. Our findings verify that the acoustic variances, recording conditions, and cultural variations in expressing emotions among sets of data pose a wide generalization gap. The fact that the model does not accurately categorize instances of No Stress in EMODB and RAVDESS means that the decision boundary trained on CREMA-D is overly aggressive with respect to datasets with varying class distributions and emotional intensive.

The lookback window ablation experiment showed a steep inflection point in performance at 10 files (around 30 seconds) with an F1-score increasing to 0.9994. This implies that 30 seconds of context is the golden mean in recording the deterministic emotional shifts in acted speech, and it is unnecessary to have further historical context.

Surprisingly, Mixup augmentation and Temporal Smoothness loss did not lead to numerical improvements on the CREMA-D validation set. This validates that the domain of acted speech is a low-noise, high-separability domain in which the underlying TCN-Attention architecture can readily memorize the patterns. Nevertheless, these regularizations, as mentioned in our methodology, are important on a theoretical level. In practice, spontaneous speech, with a background noise, speech overlap, and micro-expressions will induce smoothness in time, will make sure that flickering prediction error does not occur, and mixup will play a critical role in avoiding overfitting to speaker-specific idiosyncrasies.

This research has a number of limitations despite the good results on the acted data which define clearly the direction of future research:

1. Acted vs. Spontaneous Stress: CREMA-D data is an acted emotion data, as opposed to spontaneous and actual stress, where there is micro-expressions and physiological arousal. Future research should shift on spontaneous ecologically sound datasets to confirm the usefulness of the model in the clinical or work environments.
2. Unimodal Constraints: Although audio is sufficient to act speech, the real-world stress detection would probably need the multimodal fusion as suggested by Valerio and Mahmoud [16], who have shown that body gestures and posture are the complementary cues that cannot be provided by audio.
3. Domain Adaptation: Due to the drastic cross-dataset drop, there is need to incorporate domain adaptation methods. Future versions of this model can consider adversarial domain alignment or few-shot fine-tuning to achieve a gap between act training data and spontaneous target environment.
4. Long Term Forecasting: Although our model is effective up to 9 seconds ahead in predicting stress, clinical interventions typically need longer lead times. Combining our short-term acoustic predictions with longitudinal self-reports, e.g., the Ema2Vec time embeddings suggested by Simon and Chetouani [14], may allow multi-scale predictions of seconds to days.

## V. CONCLUSION:

This paper has examined the possibility of predicting the short-horizon stress states based on acted emotive speech using a new Multi-Task Temporal Convolutional Network with Attention (MT-TCN-Att). The major results show that the given model performs well on the CREMA-D dataset, reaching an F1-score of 0.9208 at a 9-second forecast horizon. The ablation experiments found that a lookback window of 10 files (about 30 seconds) is the best to obtain the deterministic temporal structures in acted speech and achieve near-perfect F1-scores (0.9994). Moreover, the learnable positional encoding analysis revealed that the model is based on a non-linear temporal dependency with the immediate context and the distant baseline state heavily weighed. The zero-shot cross-dataset test however revealed a strong domain shift with performance dropping to  $F1=0.6977$  on RAVDESS and  $F1=0.4246$  on EMODB, with performance heavily skewed to predicting the "Stress" category.

The study results in some important conclusions. The first is that, acoustic sequence stress prediction of the short horizon is very effective in the case of acted speech where emotional state is maintained and there is a clear difference between states. Second, a combination of the dilated causal convolution with multi-head self-attention and learnable positional encoding offers an effective architecture to learn short-range and long-range temporal dependencies in sequential affective data. Third, the stunning performance in the domain, on which the model has been trained, is contrasted by the drastic degradation in performance on unseen datasets, which

reflects that the acoustic indicators of stress which are being acquired using the acted data are not generalizing well to other corpora without domain adaptation. Lastly, temporal smoothness loss and mixup augmentation are not numerically helpful on the clean acted dataset, but are theoretically required to stabilize predictions in noisy, real-world conditions.

The paper contributes to the area of affective computing and mental health monitoring in a number of ways:

1. **Novel Problem Formulation:** As far as we know, this is the first work to specify the stress recognition as a multi-horizon sequence forecasting task (predicting 3s, 6s and 9s ahead) instead of a static classification problem, repositioning the paradigm of detecting stress responses to their active prevention.
2. **Architectural Innovation:** Our Multi-Task TCN with Learnable Positional Encoding does not only increase the accuracy of forecasts, but also has innate explainability through revealing the temporal significance of past time steps.
3. **Regularization Framework:** To enable robust deployment in ecology, we introduced a composite loss that adds temporal smoothness and employs mixup augmentation to promote logical continuity in predictions and enhance generalization.
4. **Complete Assessment:** We offered a thorough cross-dataset assessment (CREMA-D, RAVDESS, EMOdB), which constitutes a crucial reference point to comprehend the problems of domain change in acoustic stress forecasting.

Following the results, the practical measure recommendations to implement stress forecasting systems are proposed as follows:

1. **Context-Aware Deployment:** The existing system can be deployed most successfully in controlled settings (e.g., call center, simulated training) where speech patterns are similar to acted data. In the real world, it ought to be combined with domain adaptation techniques.
2. **Temporal Window Selection:** Practitioners are recommended to use a lookback window size of 30 seconds (10 files) to have adequate context to make accurate predictions, but have short forecast horizons (3-9 seconds) to act on at any moment.
3. **Clinician Interpretability:** Learned positional weights should be visualized and displayed to clinicians or users to provide an understanding as to why a stress forecast was made, increasing trust in the system.
4. **Real-World Data regularization:** Background noise and micro-expressions When switching to spontaneous speech, the loss of temporal smoothness and mixup augmentation should be turned on to avoid flickering predictions due to background noise and micro-expressions.

Although the results are promising, there are various limitations of this study:

1. **Ecological Validity:** The main training material (CREMA-D) is act portrayed emotions, not the micro-expressions and physiological arousal of the real-life spontaneous stress.
2. **Domain Generalization:** The extreme drop in performance on RAVDESS and EMOdB suggests that the model is learning dataset-specific acoustic priors, and is not learning universal stress markers, restricting its direct cross-domain application.
3. **Unimodal Constraints:** The use of audio alone disregards the complementary stress information found in facial expressions, body language and physiological indicators, which are essential to multimodal stress detection.
4. **Short Forecast Horizons:** The researchers are looking at very short horizons (not more than 9 seconds) that might not give enough lead time to execute complicated clinical procedures.

In order to overcome these drawbacks and to develop the field, the further research is to be concentrated on the following directions:

1. **Spontaneous and Ecological Data:** Future models should be trained and tested on spontaneous, in-the-wild data (e.g., recordings of everyday life) to test the forecasting architecture in ecologically legitimate contexts.
2. **Multimodal Fusion:** It is possible to combine the proposed temporal forecasting architecture with visual (facial, pose) and physiological (EDA, PPG) modalities, as Valerio and Mahmoud [16] and Ghose et al. [5] suggest, which would substantially enhance cross-dataset robustness.
3. **Domain Adaptation Methods:** Exploring adversarial domain alignment, few-shot fine-tuning, or self-supervised pre-training may contribute towards the gap in acted training data and spontaneous target settings.
4. **Long-Term Forecasting:** An extension of our short-term acoustic model to longitudinal self-report models (e.g., Ema2Vec by Simon and Chetouani [14]) would allow us to predict across multiple scales to predict just-in-time adaptive interventions (JITAs).
5. **Real-Time Edge Deployment:** To achieve real-time, privacy-preserving stress monitoring in social robotics and mobile health applications, it would be beneficial to optimize the MT-TCN-Att architecture to run on edge devices (e.g., NVIDIA Jetson) to avoid latency loss, which is the objective of Boldo et al. [1].

## REFERENCES:

- [1] Boldo, M., Forghani, D., Bombieri, N., Dautenhahn, K., & Nehaniv, C. L. (2025). A Deep Learning-Based Emotion Recognition Pipeline for Public Speaking Anxiety Detection in Social Robotics. *IEEE International Workshop on Robot and Human Communication, RO-MAN*, 1288–1293. <https://doi.org/10.1109/RO-MAN63969.2025.11217568>
- [2] Ciharova, M., Amarti, K., van Breda, W., Gevonden, M. J., Ghassemi, S., Kleiboer, A., Vinkers, C. H., Sep, M. S. C., Trofimova, S., Cooper, A. C., Peng, X., Schulte, M., Karyotaki, E., Cuijpers, P., & Riper, H. (2025). Machine-learning detection of stress severity expressed on a continuous scale using acoustic, verbal, visual, and physiological data: lessons learned. *Frontiers in Psychiatry*, 16, 1548287. <https://doi.org/10.3389/FPSYT.2025.1548287/TEXT>
- [3] Ding, D., Xu, W., Liu, X., & Zhu, T. (2025). Facial video based stress detection for enhancing ecological validity. *Acta Psychologica*, 255, 104877. <https://doi.org/10.1016/J.ACTPSY.2025.104877>
- [4] Fechtelpeiter, J., Rauschenberg, C., Goetzl, C., Hiller, S., Wierzba, E., Emonds, N., Krumm, S., Reininghaus, U., Durstewitz, D., & Koppe, G. (2025). Computational network models for forecasting and control of mental health trajectories in digital applications. *Npj Digital Medicine*, 9(1), 70-. <https://doi.org/10.1038/S41746-025-02252-3;SUBJMETA>
- [5] Ghose, D., Gitelson, O., & Scassellati, B. (2024). Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data. *ACM International Conference Proceeding Series*, 1, 22–32. <https://doi.org/10.1145/3678957.3685717>
- [6] Jeon, T., Byeol Bae, H., & Lee, S. (2024). Multimodal Stress Recognition Using a Multimodal Neglecting Mask Module. *IEEE Access*, 12, 144774–144787. <https://doi.org/10.1109/ACCESS.2024.3469575>
- [7] Kasnesis, P., Chatzigeorgiou, C., Feidakis, M., Gutiérrez, Á., & Patrikakis, C. Z. (2025). TranSenseFusers: A temporal CNN-Transformer neural network family for explainable PPG-based stress detection. *Biomedical Signal Processing and Control*, 102, 107248. <https://doi.org/10.1016/J.BSPC.2024.107248>
- [8] Kaur, R., Gitanjali, Awasthi, P. K., Dewangan, H., & Mehta, S. (2025). DeepStress: CNN-GRU-Based Workplace Stress Detection using Facial Expressions and Body Posture. *2025 International Conference on Networks and Cryptology, NETCRYPT 2025*, 811–815. <https://doi.org/10.1109/NETCRYPT65877.2025.11102191>
- [9] Kumar, A., Shaun, M. A., & Chaurasia, B. K. (2024). Identification of psychological stress from speech signal using deep learning algorithm. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, 9, 100707. <https://doi.org/10.1016/J.PRIME.2024.100707>
- [10] Lall, V., & Liu, Y. (2026). Dynamic Stress Detection: A Study of Temporal Progression Modelling of Stress in Speech. <https://doi.org/10.1109/CogMI67134.2025.00023>
- [11] Lee, C.-Y., Pleva, M., Hládek, D., & Su, M.-H. (2025). Multimodal Approaches for Stress Recognition: A Comparative Study Using the StressID Dataset (pp. 29–34). <https://aclanthology.org/2025.rocling-main.4/>
- [12] Rupp, L. H., Kumar, A., Sadeghi, M., Schindler-Gmelch, L., Keinert, M., Eskofier, B. M., & Berking, M. (2025). Stress can be detected during emotion-evoking smartphone use: a pilot study using machine learning. *Frontiers in Digital Health*, 7, 1578917. <https://doi.org/10.3389/FDGTH.2025.1578917/TEXT>
- [13] Shah, V., Patel, S., Patel, J., Ray, D., Garg, D., & Goel, P. (2025). Human Stress Detection Through Deep Learning Based Facial Expression Analysis. *2025 International Conference on Engineering Innovations and Technologies, ICoEIT 2025*, 1441–1446. <https://doi.org/10.1109/ICOEIT63558.2025.11211704>
- [14] Simon, L., & Chetouani, M. (2025). Learning Longitudinal Stress Dynamics from Irregular Self-Reports via Time Embeddings. <https://arxiv.org/pdf/2509.01569>
- [15] Staš, J., Ondáš, S., & Juhár, J. (2025). Performance Evaluation of Different Speech-Based Emotional Stress Level Detection Approaches. *IEEE Access*, 13, 112880–112904. <https://doi.org/10.1109/ACCESS.2025.3584534>
- [16] Valerio, R., & Mahmoud, M. (2025). A multimodal Framework for exploring behavioural cues for automatic Stress Detection. *ICMI 2025 - Proceedings of the 27th International Conference on Multimodal Interaction*, 1, 535–539. <https://doi.org/10.1145/3716553.3750797;CTYPE:STRING:BOOK>
- [17] ZAINAL, N. A., ASNAWI, A. L., IBRAHIM, S. N., AZMIN, N. F. M., HARUM, N., & ZIN, N. M. (2025). Utilizing MFCCs and TEO-MFCCs to Classify Stress in Females Using SSNNA. *IIUM Engineering Journal*, 26(1), 324–335. <https://doi.org/10.31436/IIUMEJ.V26I1.3411>