

# Psychophysiological Arousal in Young Children Who Stutter: An Interpretable AI Approach

HARSHIT SHARMA, Syracuse University, USA

YI XIAO, Syracuse University, USA

VICTORIA TUMANOVA, Syracuse University, USA

ASIF SALEKIN, Syracuse University, USA

The presented first-of-its-kind study effectively identifies and visualizes the second-by-second pattern differences in the physiological arousal of preschool-age children who do stutter (CWS) and who do not stutter (CWNS) while speaking perceptually fluently in two challenging conditions: speaking in stressful situations and narration. The first condition may affect children's speech due to high arousal; the latter introduces linguistic, cognitive, and communicative demands on speakers. We collected physiological parameters data from 70 children in the two target conditions. First, we adopt a novel modality-wise multiple-instance-learning (MI-MIL) approach to classify CWS vs. CWNS in different conditions effectively. The evaluation of this classifier addresses four critical research questions that align with state-of-the-art speech science studies' interests. Later, we leverage SHAP classifier interpretations to visualize the salient, fine-grain, and temporal physiological parameters unique to CWS at the population/group-level and personalized-level. While group-level identification of distinct patterns would enhance our understanding of stuttering etiology and development, the personalized-level identification would enable remote, continuous, and real-time assessment of stuttering children's physiological arousal, which may lead to personalized, just-in-time interventions, resulting in an improvement in speech fluency. The presented MI-MIL approach is novel, generalizable to different domains, and real-time executable. Finally, comprehensive evaluations are done on multiple datasets, presented framework, and several baselines that identified notable insights on CWSs' physiological arousal during speech production.

CCS Concepts: • Human-centered computing → *Ubiquitous computing; Ubiquitous and mobile computing design and evaluation methods.*

Additional Key Words and Phrases: Arousal Detection, Affective Computing, Multiple Instance Learning, Explainable AI, Machine Learning, Multi-modal Fusion, Sensors, Deep Learning, Stuttering, Children Who Stutter.

## ACM Reference Format:

Harshit Sharma, Yi Xiao, Victoria Tumanova, and Asif Salekin. 2022. Psychophysiological Arousal in Young Children Who Stutter: An Interpretable AI Approach. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 137 (September 2022), 32 pages. <https://doi.org/10.1145/3550326>

## 1 INTRODUCTION

The recent advancement in technology has resulted in the production of state-of-the-art sensors which provide an accurate reading of various physiological signals with minimum intrusion and only pose minor limitations in a person's mobility. The physiological data collected by these sensors give insight into the human affective states

---

Asif Salekin is the corresponding author.

Authors' addresses: Harshit Sharma, Syracuse University, Syracuse, USA, hsharm04@syr.edu; Yi Xiao, Syracuse University, Syracuse, USA, yxiao54@syr.edu; Victoria Tumanova, Syracuse University, Syracuse, USA, vtumanov@syr.edu; Asif Salekin, Syracuse University, Syracuse, USA, asalekin@syr.edu.

---

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2022 Copyright held by the owner/author(s).

2474-9567/2022/9-ART137

<https://doi.org/10.1145/3550326>

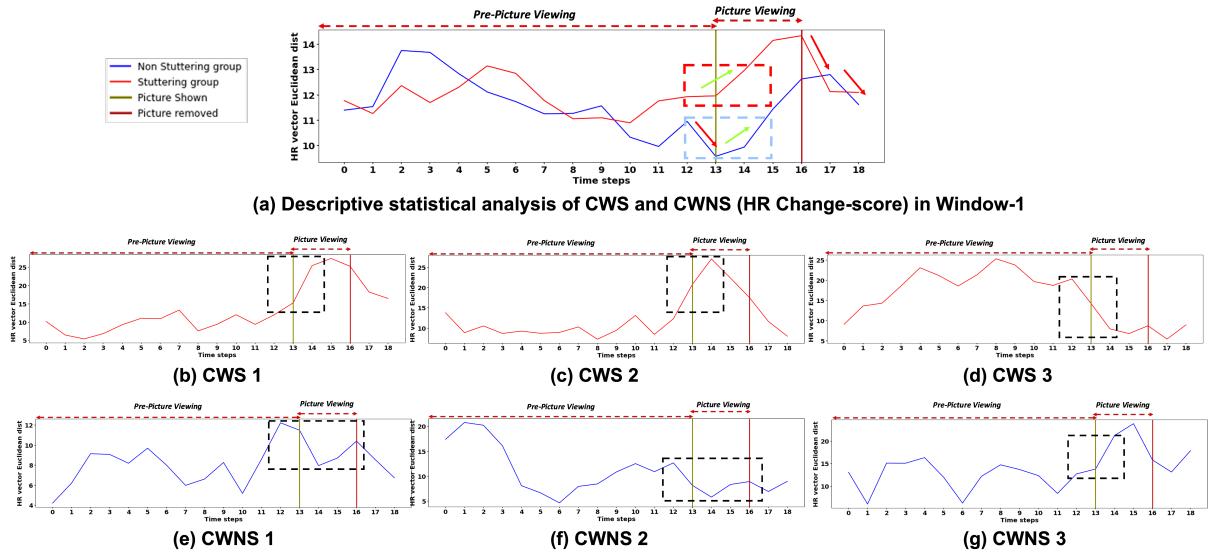


Fig. 1. (a) Shows the descriptive statistics of HR change-score Euclidean-distance-feature for CWS and CWNS in window 1 of the scripted dataset (section 2.3.2), which captures the participants' physiological response while waiting for a picture, watching the negatively-valenced picture, preparing to speak in stressful conditions. (b), (c), and (d) are three 20s physiological response (HR change-score Euclidean-distance-feature) examples from three different CWS participants. Similarly, (e), (f), and (g) are three examples from three different CWNS participants. In all figures, the dotted rectangles mark the CWS or CWNS indicative distinctive patterns present in the respective 20sec window.

and allow us to examine how emotion influences human thought and behavior. *The goal of this study is to develop automated machine learning (ML) classifiers that can identify subtle differences in affective states between young children who do stutter (CWS) and who do not stutter (CWNS) during a stressful scripted speech and a narration task.*

Emotions (i.e., affective state) are temporary, last for a short time, and are complex psychophysiological constructs composed of two underlying dimensions: valence and arousal [46, 49]. Valence is defined as the positive to negative evaluation of the subjectively experienced state [31]. Arousal measures the intensity of the affective state ranging from calm to highly excited or alert [13, 47].

Psychological arousal of an individual is observed as spontaneous responses in physiology due to an external (seeing a scary picture) or an internal (one's own thought) stressor [50]. These responses are spontaneous and can manifest themselves as changes in heart rate [52], electrodermal activity (skin conductance) [54], etc. The autonomic nervous system (ANS) is responsible for directing these physiological responses. The sympathetic nervous system (SNS), one branch of the ANS, directs the 'fight or flight' response. It stimulates the body to respond to a stressful situation by the elevation of physiological parameters like heart rate, respiration rate, blood glucose levels. The other branch of the ANS, the parasympathetic nervous system (PNS), directs the 'rest and digest' response. It conserves the body's natural activity and relaxes the individual once a stressful situation has passed. The PNS leads to a decreased arousal by reducing the heart rate and respiration rate.

The two branches of the ANS interact to coordinate our physiological responses. The interaction of the SNS and PNS, the two branches of the ANS, is demonstrated in children's cardiovascular response to a stressful situation shown in Figure 1(a). It is a descriptive statistical comparison of 'mean heart rate (HR) change-score Euclidean-distance-feature' of the CWS and CWNS participants before, during, and after seeing a negatively-valenced picture. For CWS participants, while seeing the picture, SNS causes the mean HR to increase. That is

marked by a green arrow. Once the stressful situation is over (i.e., removal of the picture), the PNS reduces the mean HR back to the resting state. That is marked by a red arrow. However, the mean response of the CWNS population follows a different trend. While exposed to an external stressor (i.e., negatively valenced picture viewing), CWNS's PNS reduces the mean HR even lower (marked by a red arrow), and subsequently, the SNS causes the mean HR to increase (marked by a green arrow). Such a response is called *freezing response* [5], which is a paradoxical decrease in HR during stressful situations. Eventually, when the stressful situation is over (i.e., removal of the picture), the PNS reduces the mean HR back to the resting state (marked by a red arrow). Notably, not all participant's physiological response follow their corresponding group's (CWS or CWNS) trend. For example, in figure 1(d), the 'CWS 3' participant's HR reduces while seeing the negatively-valenced picture, and in figure 1(g), the 'CWNS 3' participant does not experience freezing response, rather HR increases while seeing the picture; these responses are different compared to the mean response from the CWS and CWNS population, respectively. Traditional statistical methods describe the general trend of various physiological parameters as they operate on averages (e.g., calculation of the mean across the specific condition). However, they fail to identify the personalized and group-wise fine-grain, second-by-second differences (i.e., distinctive patterns) in physiological parameters between CWS and CWNS. Our observation and discussion above motivate the need for such analysis that the presented study performs.

Speech production is a complex process which requires precise coordination vocal tract while simultaneously processing cognitive-linguistic information. Social engagement, including regulating own emotions and responding appropriately to one's communicative partner, is also inherent to spoken communication. Naturally, speech production can be affected by the speaker's physiological arousal. Studies [7, 18, 36, 42, 72, 89, 90] have shown that young children who stutter are especially vulnerable to such influences.

Moreover, stuttering is a neurodevelopmental speech disorder [70] that emerges in early childhood (between the ages of 2 and 4), hence, it is essential to examine the effects of physiological arousal on speech characteristics in young children as opposed to adults. Given that preschool-age is the time when essential communication skills are undergoing most significant development and also when some children develop stuttering, it is essential for our understanding of stuttering to examine young children's physiological response during speech production.

*This study presents an interpretable AI approach to identify the second by second fluctuations and pattern differences in physiological arousal of preschool-age children who stutter compared to others who don't during various speaking tasks.* This level of analysis could inform our understanding of how emotional arousal could contribute to the development of stuttering overall and whether subtle differences in physiological arousal can lead to immediate changes in speech fluency/disfluency and speech articulation.

Moreover, the developed machine learning classifiers identify personalized distinctive situational physiological arousal for each 20sec physiological sensing data. Hence, they can be leveraged for remote, continuous, and real-time assessment of CWS's physiological arousal and may lead to automated, personalized, and just-in-time interventions to mitigate their physiological arousal, consequently mitigating stuttering disfluency.

*What follows is an overview of the nature of stuttering, the autonomic nervous system activity in response to speaking in CWS and the potential role of physiological reactivity in speaking and the development of stuttering (section 1.1). Further, we discuss the deep learning frameworks to understanding of children's physiological response during speech (sections 1.2, and 1.3). We end the introduction with the study challenges (section 1.4), and our research questions, problem statement and contributions (section 1.5).*

## 1.1 Physiological Arousal in Children who Stutter

Stuttering is believed to be a multifactorial condition where multiple factors interact and contribute to stuttering onset in early childhood and its later development to its chronic form [70]. Research shows that speech production leads to increased autonomic arousal in both adults [33, 41, 83], and children [7, 18, 36, 42, 72, 89, 90]. In this

study we were interested in examining whether speaking tasks that vary in linguistic complexity are inherently more stressful and associated with higher physiologic arousal for preschool-age children who stutter. We were also interested in examining any patterns in physiological reactivity that distinguish children who stutter from their typically fluent peers. Psychophysiological research to date offered mixed findings regarding whether preschool-age children who stutter differ in their autonomic arousal during speech production from their peers who do not stutter.

Literature from speech science generally indicates that preschool-age children who stutter do not have an elevated autonomic arousal during such speaking tasks as picture naming, picture description, and non-word repetition [7, 18, 36, 37, 42, 72, 78, 79, 89, 90]. However, differences in autonomic arousal between children who do and do not stutter based on the children's age [89] and complexity of the speaking task [72] have been reported. Further, among children who stutter, differences related to children's stuttering chronicity [90], and speech fluency [79] have also been observed. Importantly, studies published to date relied on traditional statistical approaches in examining potential differences in physiological arousal between children who do and do not stutter. However, identifying fine-grain, second-by-second, and personalized differences in physiological parameters between CWS and CWNS groups during speaking tasks has yet to be addressed.

## 1.2 Deep Learning in Physiological Arousal and Stress Detection

Various ML approaches (e.g., SVM, CNN, RNN, RCNN) have been developed to detect arousal/stress through physiological sensors [4, 34, 53] from non-stuttering individuals. For example, a study [61] evaluated a convolutional neural network (CNN) and long short-term memory (LSTM), taking a combination of ECG signal features, vehicle, and contextual data as input to predict driver stress with an accuracy of 92.8%. Another study [4, 53] used CNNs and Recurrent CNNs to detect the stress using physiological parameters such as electrodermal activity (EDA) as features and achieved an accuracy of 67.50% and F1-score of 0.71 respectively in detecting stress. Table 10 in Appendix 8.6 summarizes the recent literature on physiological sensing-based stress assessment.

## 1.3 Deep Learning and Physiological Sensing for Stuttering Speech Disfluency detection

To the best of our knowledge, no study has developed ML classification approaches to differentiate the physiological responses between CWS vs. CWNS during perceptually fluent speaking tasks like scripted phrase repetition under an external stressor and spontaneous narration conditions. However, a study [77] on adults with stuttering (AWS) has developed a multi-layered perceptron (MLP) neural network-based disfluency classification system. It used respiration signals as input and achieved an 82.6% accuracy in differentiating the physiological response of 'AWS during disfluent speech production' vs. 'AWNS during fluent speech production.' In this study [77], the participants were asked to read aloud a 677-syllable text extracted from a Spanish story. The speech-language pathologists observed the experiments and provided fine-grain annotations of stuttering disfluency events of the AWS, such as sound/syllable repetitions, sound prolongations or silent tense pauses. *Notably, in contrast, in our data (section 2), children are talking perceptually fluently for approximately 97% of the duration. Hence, the presented study focuses on differentiating the physiology of the CWS vs. CWNS during both perceptually fluent and disfluent speech production.* Though statistical-based studies (discussed in section 1.1) have shown that there are group-wise differences in the physiological response of CWS vs. CWNS during perceptually fluent speech production, no study to our knowledge has examined or identified the second by second salient and distinguishable patterns present in CWSs' vs. CWNSs' physiology during their fluent speech. Since such patterns are unknown, it is not possible to annotate our data at the fine-grain level, which makes our classification a challenging task [30, 85].

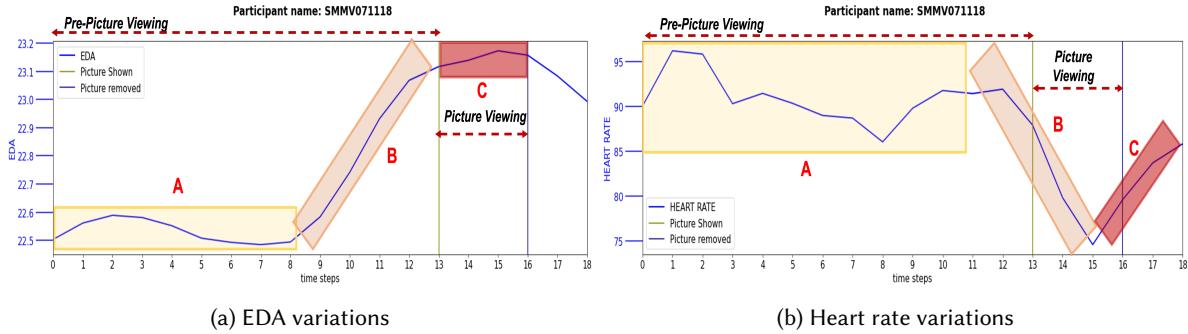


Fig. 2. EDA and HR signals during negatively valenced picture viewing for a CWS participant. The 'Yellow' regions (i.e., region A) of the signals do not convey any arousal indicative pattern. 'Orange' regions (i.e., region B) show the signal patterns indicative of low-to-high arousal transition, and the 'red' regions (i.e., regions C) show the high arousal response.

## 1.4 Study Challenges

This paper aims to develop automated ML classifiers that can identify the subtle differences in affective states between CWS and CWNS during the stressful scripted speech and narration tasks. This section discusses the challenges of developing such a classifier effectively.

**1.4.1 Weakly Labeled Data:** A challenge is, though prior statistical studies [72, 89] established the existence of physiological parameter differences in CWS and CWNS, none of them identified the CWSs' distinctive physiological signal patterns. Moreover, preschool-age children cannot self-assess and report on their physiological state due to their young age [32]. Also, changes in arousal would not always be accompanied by observable changes in behavior. Ergo, our data is ‘weakly labeled’, meaning we do not have the precise labels of CWS’s distinctive physiological signal patterns. For example, figure 1(b)-(g) show six children’s 20s cardiovascular response (i.e., HR change-score) window. The dotted rectangles show the children’s class (CWS or CWNS) indicative distinctive patterns. These patterns are 2 – 5s in duration and may appear in any timestamps of the 20s window. Our data annotations only provide information about the data belonging to a CWS or CWNS individual; the distinctive patterns and their timestamps in the detection window are unknown. Additionally, our datasets are limited in size. There were 180 and 200 picture-viewing events for CWS and CWNS during the scripted speech experiment. Supervised learning classifiers fail to learn the above discussed subtle, sparse and independent physiological sensing patterns from ‘weakly’ labeled data, specifically while limited in size.

**1.4.2 Modality-wise Distinctive Patterns:** This paper evaluates multiple physiological modalities (e.g., HR, EDA and Respiration activities) to measure the participants' (i.e., CWSs' and CWNSs') physiological response. A notable observation is that arousal indicative sparse patterns do not simultaneously emerge in each modality. Figure 2 shows the EDA and HR signals of a CWS participant during a negatively valenced picture viewing. As shown in Figure 2a, a low-to-high arousal transition (i.e., region B) appears in the EDA during 8s to 13s timestamps while the participant is waiting for a picture to be viewed (anticipation effect [28]), and the participant is experiencing high arousal during the negatively valenced picture viewing (i.e., region C). However, according to Figure 2b, a low-to-high arousal transition (i.e., region B) appears in the HR during picture viewing (12s-15s timestamps), where the HR decreases rapidly (*freezing response* [5]). High arousal (i.e., region C) in HR is depicted as the subsequent increase in HR after the freezing response. Notably, our 'weakly' labeled data do not have the fine-grain annotations of such arousal indicative of sparse patterns in any of the modalities. The above-discussed

observation motivates that *our developed classifier needs to extract CWS's distinctive sparse patterns independently from each modality, without any available annotations (for such patterns) during training.*

**1.4.3 Capturing Cross-modality Dependency:** Previous studies [10, 40] have shown that the correlations between physiological parameters such as HR, EDA and Respiration activities are effective attributes in stress [4, 34, 53] or emotional valence [68] detection. Moreover, the HR-EDA synchronization is positively associated with variability of arousal responses [40]. Hence, our solution approach must capture the cross-modality relationships as well.

*In conclusion, our solution needs to identify CWS's distinctive sparse patterns independently from each physiological modality without any available annotations of such patterns, and capture and leverage the cross-relationships of the identified modality-specific sparse patterns for effective CWS vs. CWNS classification.*

## 1.5 Problem Statement and Contributions

In particular, the study aims at investigating the following key questions. To our knowledge, no previous study has investigated these questions through machine learning approaches.

**(1) Do the CWS and CWNS show different physiological responses to external stressors?**

This question aims to find differences in the physiology of the CWS and CWNS group under the arousal inducing condition when the participants are viewing negatively-valenced pictures. To investigate, we developed and evaluated classifiers to differentiate the CWS vs. CWNS from the 20s 'window-1' signals of the scripted dataset (discussed in sections 4.1.1 & 2.2).

**(2) Do the CWS and CWNS show different physiological responses while perceptually fluently talking under stressful conditions?**

This research question aims to understand the differences (i.e., if any) in the physiology of the CWS vs. CWNS groups while talking under stressful conditions (i.e., after viewing negatively-valenced pictures). To investigate, we developed and evaluated classifiers to differentiate the CWS vs. CWNS from the 20s 'window-2' signals of the scripted dataset (discussed in sections 4.1.2 & 2.2).

**(3) Do the CWS and CWNS show differences in physiology under rest or baseline condition?**

To investigate, we developed and evaluated classifiers to differentiate the CWS vs. CWNS from the 20s baseline/neutral condition physiological signals of the scripted dataset. (discussed in section 4.1.3)

**(4) Do the CWS and CWNS show different physiological responses during spontaneous narration?**

The narration task is linguistically and cognitively demanding since the children develop new context or storylines and articulate them in speech. Therefore, this research question investigates whether the CWS show different physiological responses than the CWNS while performing such narration task. To investigate, we developed and evaluated classifiers to differentiate the CWS vs. CWNS from the 20s windows of narration task condition of the free-speech dataset (discussed in section 4.1.4).

**1.5.1 Contribution in Classification:** To investigate the above-mentioned questions, we develop and evaluate a *novel Modality Invariant-MIL (MI-MIL) (section 3.2) classifier.*

- (1) To address the weakly labeled data challenge and identify modality-wise distinct patterns (section 1.4), *MI-MIL applies modality-wise multiple-instance-learning (MIL) paradigm in each physiological modality independently.* MIL paradigm is designed to extract sparse and subtle patterns from weakly labeled data (i.e., without any fine-grain annotations of the region, timestamps, or duration of the patterns in the data).
- (2) To capture the cross-modality-relations (section 1.4), *MI-MIL presents a novel modality-fusion network that identifies the cross-relations of each modality's CWS indicative sparse patterns.*
- (3) Our evaluations discussed in section 4.1 show, the presented approach outperforms the supervised learning classifiers, and the recent state-of-the-art MIL approach (attention-based MIL) significantly.

- (4) Our evaluation demonstrates, MI-MIL is real-time executable in scalable resource constraint devices: NVIDIA Jetson Nano and Google pixel 6 smartphone.

Developed MI-MIL models' high efficacy in addressing each of the four questions (*q1-q4*) indicates the existence of physiological signal patterns that differentiate the binary categories, and our classifier can identify them.

### 1.5.2 Contributions in Dataset Collection and Physiological Features:

- (1) We collected *two datasets* (section 2) containing CWS and CWNS's *physiological responses* (i.e., physiological sensing parameters: electrodermal activity - EDA, heart rate - HR, respiratory rate - RSP-rate, and respiratory amplitude - RSP-amp) while speaking in two challenging conditions: speaking under stressful conditions (after experiencing external stressors) and a linguistically and cognitively demanding narration task where the children need to spontaneously develop new context and concepts.
- (2) Motivated by the state-of-the-art behavioral science studies, we extracted *a novel vector-distance-based representation of change-scores features* of physiological parameters that captures the fluctuation of one's physiological response in a target condition than their neutral condition (section 2.3.2). Presented change-score representation significantly outperforms the representation used in literature (Appendix 8.1).
- (3) We evaluated ML models for both raw and change-score physiological features. They give us insights into the CWS vs. CWNSs' physiological-patterns and fluctuations-differences. Furthermore, we evaluated and ranked the raw features according to their discriminative capabilities (section 4.2). The feature importance ranking is in line with our observation of Shapley visualization (section 5).

**1.5.3 Classification Interpretation:** This study is the first-of-its-kind to analyze, interpret, visualize and discuss the fine-grain, second by second, temporal, and distinctive physiological response patterns of CWS from CWNS during speech production in different challenging conditions. The developed MI-MIL classifiers are black-box models; hence we employ SHAP [16] ML model explainers (detailed discussion is in section 5) to extract and visualize the *group-wise* and *personalized* situational physiological arousal patterns. Identifying and visualizing group-wise patterns would enhance our understanding of stuttering etiology and development. Personalized pattern identification would enable remote, continuous, and real-time assessment of stuttering children's physiological arousal, which may be used clinically to develop personalized emotion regulation strategies (e.g., biofeedback, mindfulness intervention), resulting in an improvement in speech fluency.

## 2 DESCRIPTION OF THE DATASET AND OUR DATA COLLECTION PROCEDURE

We collected two datasets comprising preschool-age children's (both CWS and CWNS) physiological response (i.e., physiological sensing parameters) while performing two different speaking tasks. Due to the nature of speaking tasks, we refer to the first dataset as "free speech dataset" and to the second dataset as "scripted dataset" consistently throughout the manuscript. Study participants and procedures are explained below.

**Participants:** Participants in both datasets were preschool-age children (mean age: 50.3 months, std: 9.14). The study procedures were approved by the Syracuse University Institutional Review Board. All data collection procedures took place in our Laboratory over two visits. During the first visit participants were administered standardized tests of speech and language to ensure age-appropriate speech articulation and language scores, and passed a pure-tone hearing screening. Participants' speech fluency was assessed by a licensed speech-language pathologist and the diagnosis of stuttering was established using evidence-based diagnostic criteria [73, 88]. All psychophysiological data were collected during the second visit. The free-speech dataset comprised data from 35 preschool-age children (age range: 36-67 months). Among the participants, 16 were CWS, (13 boys and 3 girls), and 19 were CWNS (12 boys and 7 girls). The scripted dataset comprised data from 38 preschool-age children (age range: 38-69 months). Among the participants, 18 were CWS, (16 boys and 2 girls) and 20 were CWNS (16 boys and 4 girls).

## 2.1 Data-Collection Experimental Procedure

We explain the common procedures followed in data collection for both datasets first, then we describe the dataset specific differences in procedures. Upon arrival at the lab, participants played and spoke with the examiner for about *15 minutes* to get them acquainted with the lab. Then, they were seated in a child-sized chair, in front of a computer screen. Hypoallergenic electrodes were attached to the skin at the suprasternal notch of the rib cage and at the 12th rib laterally to the left for acquisition of the electrocardiogram [75]. A strain gauge transducer designed to measure respiratory-induced changes in thoracic or abdominal circumference (model TSD201, Biopac Systems, Inc.) was used to record respiratory effort. The transducer was positioned around the participants's chest for acquisition of the respiration waveform. The electrodermal activity was recorded with electrodermal response transducers (model TSD 203, Biopac Systems, Inc.) which included a set of two Ag-AgCl electrodes with incorporated molded housings designed for finger attachment. The response transducers were filled with an isotonic electrolyte gel and were placed on the volar surfaces of the middle phalanges of the two fingers of the participants' right hand. After the sensors were placed, the participants' baseline psychophysiological data were collected first followed by the experimental conditions. The conditions are explained below.

**2.1.1 Baseline Condition:** For both datasets, to establish a pre-experimental baseline for each participant's resting skin conductance level, breathing rate and heart rate, participants viewed an animated screensaver of a three-dimensional fish tank for *four minutes*. This procedure has been successfully implemented in prior studies to establish baseline psychophysiological levels in preschool-age children [36, 72].

**2.1.2 Experimental Condition - Free Speech Dataset:** The experimental condition in this dataset was a picture description task, which lasted about *10 minutes*. During the picture description task, participants were shown pictures from a wordless storybook about a boy, a dog, and a frog by the author *Mercer Mayer, Frog Goes to Dinner* [55]. To keep the narrative elicitation procedure consistent between the participants, the examiner was not allowed to ask specific questions about the picture but could only prompt the participant to tell them what was happening in the picture by saying "Let's look at this picture. Tell me what is happening here." The examiner was instructed to provide no more than three such elicitation prompts per picture. Participants who stuttered did show some (3% of the total speech) stuttering events, such as sound repetitions and prolongations, during this experiment.

**2.1.3 Experimental Condition - Scripted Dataset:** The experimental condition in this dataset also lasted approximately *10 minutes* and involved negatively-valenced picture viewing and phrase repetition. Specifically, the participants were shown 10 negatively-valenced pictures from the International Affective Picture System[48] and were asked to repeat a target phrase "Buy Bobby a puppy" (BBAP phrase) after a pre-recorded prompt presented over the speakers. Picture presentations were interspersed with speaking such that per each picture shown the participants were asked to repeat the target phrase 3 times. Figure 3 shows the chronological order of events for one negatively-valenced picture viewing in the scripted dataset. None of the participants showed any stuttering disfluency events during this experiment.

**2.1.4 Data Acquisition: Preprocessing and Cleaning:** The respiratory, electrodermal, and cardiac activity were acquired simultaneously using the Biopac MP150 hardware system (Biopac Systems, Inc.) and recorded using Acqknowledge software (ver. 4.3 for PC, Biopac). Respiratory effort (RSP), electrodermal activity (EDA), and electrocardiogram (ECG) signals were sampled at 1250 Hz during the baseline and experimental conditions. The butterworth high pass filter[62] was applied to the raw signals to remove the noise and baseline drift.

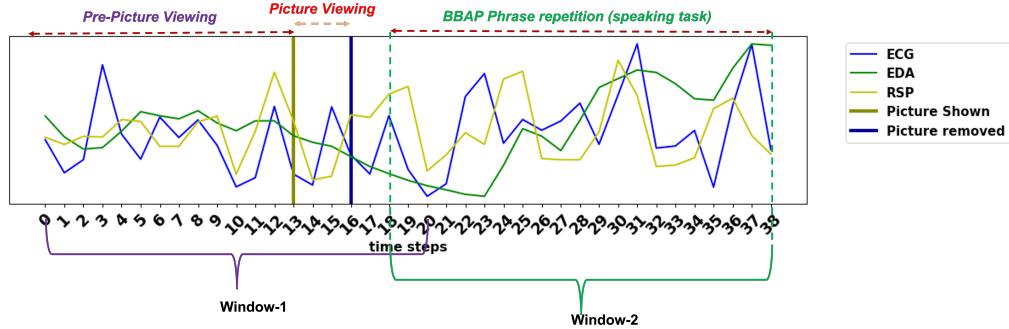


Fig. 3. Chronological order of events for one negatively-valenced picture viewing in the scripted dataset.

## 2.2 Extraction of Event Detection Windows:

This paper detects the affective state differences from 20s windows; since in the scripted dataset, children took on avg. 20s to utter BBAP phrase 3 times following a picture viewing. For the *free-speech dataset*, baseline-level data collection and picture description sessions lasted approximately 4min and 10min. We segment the sessions into 20s windows with a 5s overlap (15s hop-length). Physiological sensory streams from each 20s window from baseline and picture description session represent the participant's physiological response in the neutral and narration (i.e., linguistic and cognitively demanding) conditions. Similarly, in the *scripted dataset*, baseline-level data collection lasted approximately 4min, which was segmented into 20s windows. Sensory signals from these 20s windows represent participants' physiological response in the neutral condition. During the speaking task, for each negatively-valenced picture viewing, we extracted two 20s windows: (1) during picture viewing; and (2) after the picture viewing while the participants were repeatedly saying the scripted phrase ("Buy Bobby a puppy"). As shown in figure 3, the picture was flashed on the screen for 3 seconds (i.e., red region), after which the computer prompts the sentence "Buy Bobby a puppy" (BBAP). After that, participants started to speak, which took approximately 20s. The first window starts at time step 0 (where in each time step is a 2s window) and the image is shown at time step 13. Hence, window 1 of the scripted dataset captures the participants' physiological response while waiting for a picture, watching the pictures, preparing speech in stressful conditions. That is, it captures the participants' progression from normal to stress state. Window 2 started at time-step 18 and ended at time step 38. Hence, Window 2 of the scripted dataset captures the participants' physiological responses while speaking in stressful conditions. (It is important to note that window-1 and window-2 both consist of nineteen time-steps or 2s instance windows each have 1s overlap resulting in 19 time-steps for a 20s window). The data collection sessions were segmented into 20s windows with a 5s overlap; hence for the free speech dataset, we had a total of 1510 windows. For the scripted dataset, there were 752 windows in total; wherein there were 376 window-1 and window-2 each.

## 2.3 Physiological Features Extraction

As discussed in the previous section each of the 20s signals was divided into nineteen segments with 2s duration and 1s overlap, and features are extracted from the 2s segments. We evaluated the event detection models using two categories of features: (a) raw features and (b) change-score features.

**2.3.1 Raw Physiological Features.** Psycho-physiological features relevant to affective states and stuttering individual's physiological responses are extracted from each 2s segments. We extracted the *heart rate (HR)* from the raw ECG signal, since a recent study [74] of heart rate (HR) in relation to stressful situations indicates that children who stutter showed a significantly higher HR than CWNS. *Electrodermal activity (EDA)* signals increases and shows spontaneous fluctuations during arousal [9, 58]. Hence, filtered EDA is one of the extracted features.

Moreover, studies have shown respiratory distributions vary in individuals with stuttering vs. non-stuttering and different speech-fluency levels [71]. To capture the children's respiratory patterns, we extract the *respiratory rate (RSP-rate)* and *respiratory amplitude (RSP-amp)*[66] as raw physiological features from the raw RSP signal.

Following previous studies [63, 69], two-level physiological features (Low-level and high-level Descriptors) are extracted, allowing the ML models to capture signal characteristics in different granularity levels. Four low-level descriptors (LLD) features: heart rate (HR), electrodermal activity (EDA), respiratory rate (RSP-rate), and respiratory amplitude (RSP-amp), were extracted at 0.8-millisecond intervals from each 2s segment. Six high-level descriptors (HLD) functionals: min, max, std, var, mean, and median, are applied to the LLDs to extract the feature representation for each of the four LLDs, totaling <24> HLD features, are extracted from each 2s segment. In total, we extract  $19 \times 24$  raw physiological features from the 20s event detection window.

**2.3.2 Change-score Features.** State-of-the-art behavioral science studies evaluate change-score [17, 19, 37] features to understand the psychophysiological changes on individuals in response to different affective states (e.g., arousal). A change score is the difference between the value of a variable/feature measured at one point in time ( $Y_t$ ) from the average value of the variable for the same unit at the baseline-level condition ( $Y_b$ ).  $Y_t$  is called the 'post-score,'  $Y_b$  is the 'baseline-score,' and the difference between  $Y_t$  &  $Y_b$  is the 'change-score'. This study extracts change-scores of HR, EDA, RSP-amp, and RSP-rate LLD-features from each 2s segment. The post-, baseline-, and change-scores of these physiological features are represented as vectors.

*Post-scores.* are calculated from each 2s segments in different non-baseline scenarios. Each of the LLD features' is represented as a 6-dimensional vector (i.e., one dimension for each HLD-functional) quantifying an individual's physiological response. These 6-dimensional vectors are the respective LLD features' post-scores.

*Baseline-scores.* are calculated from all of the 2s segments in the individual's baseline condition. For each LLD feature (HR, EDA, RSP-amp, and RSP-rate), we consider the mean of its 6-dimensional HLD vectors from all baseline-condition 2s segments as its 'baseline-score' vector. Meaning, we extract four 6-dimensional baseline-scores/vectors for each individual, representing the average LLD-features values in their neural condition.

*Change-scores.* are the vector differences between the post-scores and baseline-score. For each 2s non-baseline segment, it quantifies the difference in an individual's physiological response regarding their baseline (i.e., neutral) response.

In this study, the vector difference between the post-score and baseline-score is measured by two matrices: cosine similarity and the euclidean distance. They are the two most common matrices to measure vector difference used in machine learning [26, 44, 76, 80].

For each of the four LLD features HR, EDA, RSP-amp, RSP-rate, we calculate two change-score values (euclidean distance and cosine similarity), totaling eight change-score features extracted from each 2s small-signal segments. In total, we extract  $19 \times 8$  change-score features from the 20s event detection window.

**2.3.3 Comparison of Features:** Raw physiological features (HLDs) *captures the signal amplitudes and patterns of individuals in different conditions*. Hence, affective state detection classifiers learn the physiological parameter values and time progression relevant to different target classes (e.g., arousal). However, a limitation is that individuals' physiological responses in the neutral condition can be dissimilar. Hence, a classifier trained on raw physiological signals may misclassify an individual's neutral state to aroused state if their baseline-level physiological signal attributes are different from the average population. Change-score features eliminate such bias since they *capture the difference in an individual's physiological response in different conditions compared to their neutral state*. However, in doing so, change-score features lose fine-grain information of the signals (e.g., signal amplitudes, std, etc.). Hence, *this paper trained and interpreted classifiers using both raw and change-score physiological features to understand how their 'attributes (amplitudes and patterns)' and 'fluctuations compared*

*to the neutral condition*' indicate children's mental states during phase repetition under aroused condition and narration tasks.

### 3 METHODOLOGY: CWS VS. CWNS PSYCHOPHYSIOLOGICAL AROUSAL DETECTION

#### 3.1 Approach Design Choices

This section discusses how our presented MI-MIL approach addresses the challenges (discussed in section 1.4) in CWS vs. CWNS affective states difference classification from physiological sensing signals.

**3.1.1 Weakly Labeled Data and Multiple Instance Learning (MIL):** We employ a multiple instance learning (MIL) paradigm [64, 86, 87] to address the absence of fine-grain annotations in our data. In MIL, each input of a classifier is considered as a bag of instances  $B = \{x_1, x_2, \dots, x_K\}$ . Each bag  $B$  has an associated single binary label  $Y \in \{0, 1\}$  known during training. However, the labels of instances within a bag, i.e.,  $y_1 \dots y_K$  and  $y_k \in \{0, 1\}$  for  $k = 1 \dots K$  are unknown. As per conventional instance-based MIL assumption [24], a positive bag has a label:  $Y = 0$  and a negative bag has a label:  $Y = 1$ . A negative bag has at least one negative instance, and may contain positive instances (i.e.,  $\exists x_j \in B, y_j = 1$ ). However, a positive bag contains positive instances only (i.e.,  $\forall x_j \in B, y_j = 0$ ). Thus, the relationship between bag label  $Y$  and instance label  $y_j$  is:  $Y = \max_{i=1\dots k} (y_i)$ .

In this paper, a negative bag is the extracted features set from a CWS's 20s physiological data, whereas a positive bag is from a CWNS. Features extracted from each 2s segment discussed in section 2.3 constitute an instance, and the collection of all instances of a 20s physiological sensing data constitute a bag. According to the instance-based MIL assumption, if the MIL classifier identifies that at least one instance is negative (2s segment, conveying CWSs' distinctive physiological response pattern), the 20s data would be detected as a CWS's response. In contrast, if none of the instances is identified as negative, the 20s data would be detected as a CWNS's response.

**3.1.2 Temporal Dependency and Attention-MIL:** Previous studies [8, 79] have shown that the physiological response in arousal comprises temporal patterns. A limitation of instance-based MIL [24] is that it considers that the distinctive patterns (indicative to CWS or negative class) are sparse and independent. Moreover, it classifies a bag (i.e., one input) based on the instance with the highest likelihood of being negative. Hence, a large portion of the data remains unutilized, negatively affecting the classification performance.

To address this challenge, this paper adopts an attention-based MIL approach named attention-MIL [35, 81]. Attention-MIL is capable of identifying sparse distinctive patterns from weakly labeled data, captures sequential traits, makes inferences from the aggregation of all instances in a bag, and is shown to achieve better classification performance [81]. In contrast to the instance-based MIL, it generates a score or attention weight for each instance ( $x_i$ ) in a bag indicating the likelihood of the instance ( $x_i$ ) conveying CWS indicative distinctive patterns. Weighted instances (preserving their temporal patterns) are aggregated through an attention-based pooling function (section 3.2.2) to generate a bag representation, from which the classifier makes the inference.

**3.1.3 Modality-wise Distinctive Pattern Extraction:** This paper evaluates multiple modalities: HR, EDA, RSP-amp and RSP-rate (section 2.3) to measure the participants' physiological response. Section 1.4's observation demonstrates that arousal indicative sparse patterns do not simultaneously emerge in each modality. Such observation motivates the need for CWSs' distinctive sparse patterns extraction independently from each modality. Hence, this paper applies the attention-MIL approach to each modality separately and generates modality-specific bag representations.

**3.1.4 Capturing Cross-modality Dependency:** As discussed in section 1.4, cross-modality relationships can be effective attributes in differentiating CWS's situational physiological response from CWNS. Hence, the presented approach must capture the cross-modality relationships.

As discussed above, the modality-wise attention-MIL mechanism generates an independent representation for each modality. To capture the cross-modality relations, the presented approach uses a novel *modality fusion mechanism* which learns the pair-wise and unary relationships between each modality-embedding.

### 3.2 Modality Invariant-MIL (MI-MIL) Approach

This section discusses the MI-MIL approach that takes the modality-specific bag representations ( $B_m = \{x_{1m}, x_{2m}, \dots x_{km}\}$ ,  $k = 19, m = \text{EDA, HR, RSP-amp, RSP-rate}$ ) of a 20s physiological sensing data as input. As shown in figure 4, MI-MIL has four components: (1) modality specific embedding block, (2) modality specific self-attention pooling block, (3) modality fusion Block, and (4) classifier Block. While the first two blocks are applied to each modality  $m$  independently, the latter two combine the cross-modality information to generate inference. The components are discussed in detail below.

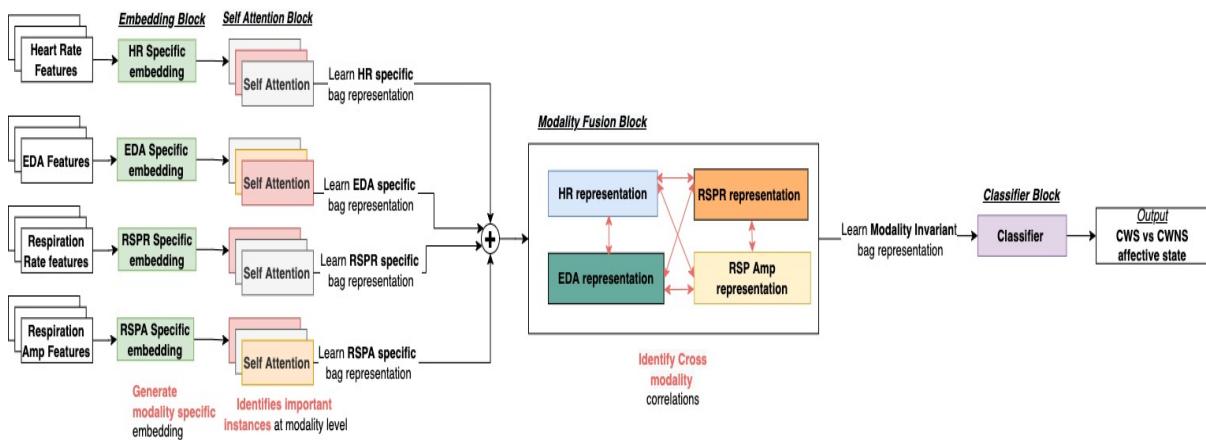


Fig. 4. Modality-Invariant Multiple Instance Learning (MI-MIL). In this example figure,  $k = 3$ , and four modalities  $m = \text{EDA, HR, RSP-amp and RSP-rate}$ . Each modality-specific input bag representation  $B_m$  comprises three instances  $x_{im}, i = 1, 2, 3$ . Modality specific embedding blocks generates three embeddings  $e_{im}, i = 1, 2, 3$ . Each self-attention pooling block generates attention weights  $a_{im}$  for the respective modality instance embeddings  $e_{im}$ . Weight values are shown with color, where darker color represents higher weight values. Each self-attention pooling block generates a modality-specific bag representation  $t_m$  using a weighted average of the modality-specific embeddings. The modality fusion block takes the  $\langle t_{EDA}, t_{HR}, t_{RSP-amp}, t_{RSP-rate} \rangle$  vector as input and generates a modality invariant representation  $Z$ , conveying the cross-modality relations. A classifier takes the  $Z$  as input and infers the class label (CWS vs. CWNS).

**3.2.1 Modality Specific Embedding Block:** For each modality  $m$ , the MI-MIL utilizes a modality-specific embedding block  $f_m$  that takes each 2s segment (i.e., modality-specific instance  $x_{im}$ ) of the respective modality as a separate input and transforms it into a lower  $p$  dimensional embedding vector ( $e_{im}$ ) [25]. Each block  $f_m$  comprises multiple linear layers with ReLU activation functions. Generated embeddings  $e_{im}, i = 1, 2, \dots, k$  convey the modality  $m$  specific CWS vs. CWNS differentiating information from their respective instance  $x_{im}$ .

**3.2.2 Modality Specific Self-Attention Pooling Block:** MI-MIL utilizes a modality specific self-attention pooling block for each modality  $m$ . It takes the modality specific embeddings  $e_{im}, i = 1, 2, \dots, k$  preserving their temporal order as input, generates attention-weights  $a_{im}$  for each of them, and generates a modality specific bag embedding

$t_m$  following the equation 1.

$$t_m = \sum_{i=1}^k a_{im} e_{im} \text{ where } a_{im} = \frac{\exp\{w_m^T \tanh(V_m e_{im}^T)\}}{\sum_{j=1}^k \exp\{w_m^T \tanh(V_m e_{jm}^T)\}} \quad (1)$$

Here,  $w_m \in R^{L \times 1}$  and  $V_m \in R^{L \times M}$  are the network parameters of the  $m$  modality-specific self-attention pooling block. The hyperbolic tangent  $\tanh(\cdot)$ , element-wise non-linearity is utilized to ensure proper gradient flow [35]. Generated weights  $a_{im}$  represent the likelihood of the embedding  $e_{im}$ 's conveying CWS vs. CWNS differentiating pattern information, and the weights  $a_{im}, i = 1, 2, \dots, k$  must sum to 1 to be invariant to the number of instances of a bag. Hence, this block ensures the temporal patterns present in each modality are captured, aggregate CWS vs. CWNS differentiating modality-specific information from each 2s segments (i.e., instances), and trainable through backpropagation.

As discussed in section 3.1, CWS vs. CWNS differentiating patterns can be present in different asynchronous portions (i.e., different timestamps) of different modality-signals. Hence, different modality-specific pooling blocks may learn different attention weights for each 2s segment (i.e., instance), enabling CWS indicative pattern extraction independently from each modality. Equation 1 is similar to the attention pooling mechanism presented in attention-MIL paper [35].

**3.2.3 Modality Fusion Block:** Modality fusion block captures the cross-modality relationships. It receives four independently generated modality-specific bag embeddings from four modality-specific self-attention pooling blocks. Each of the embeddings is an N-dimensional vector. It concatenates the four embeddings to a  $[1, 4N]$  dimensional vector,  $X = < x_1, x_2, \dots, x_{4N} >$ , and generates a  $[1, 4N]$  dimensional vector,  $Z = < z_1, z_2, \dots, z_{4N} >$  that encodes the pairwise relations among all possible vector dimensions  $(x_i, x_j)$  of  $X$ , as well as the unary relations, meaning how one dimension  $x_i$  of  $X$  may have its independent impact. Hence,  $Z$  essentially encodes pair-wise and unary relationships between each of the dimensions of each of the modalities. Each  $z_i$  encodes the relations corresponding to  $x_i$  and computed using equation 2.

$$z_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \quad (2)$$

Here,  $z_i$  computation enumerates all possible dimensions  $j$ .  $f(x_i, x_j)$  represent the *pairwise* relation between dimension  $i$  and  $j$  of  $X$ . Here we use Embedded Gaussian function as function  $f$ :  $f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)}$ .

Here,  $\theta(x_i) = W_\theta x_i$  and  $\phi(x_j) = W_\phi x_j$  are embeddings of  $x_i$  &  $x_j$ , and  $W_\theta$  &  $W_\phi$  are learnable network parameters. In our implementation,  $W_\theta$  &  $W_\phi$  are single-convolutional-layers with kernel size of 1.  $\theta(x_i)^T \phi(x_j)$  is the dot-product similarity. The normalization factor is set as  $C(x) = \sum_{\forall j} f(x_i, x_j)$ . With the equation above,  $\frac{1}{C(x)} f(x_i, x_j)$  become a softmax operation along the dimension  $j$ .

The function  $g(x_j)$  generates an unary embedding of  $x_j$ . It is a simple linear embedding:  $g(x_j) = W_g x_j$ , where  $W_g$  is a learnable network parameter. In our implementation,  $W_g$  is a single-convolutional-layer with a kernel size of 1. Hence, according to equation 2, the modality fusion block generates a modality invariant representation  $Z$ , encoding pair-wise relations between each of the modalities while preserving each modality's unary information.

**3.2.4 Classification Block:** The classification block predicts the bag label (CWS vs. CWNS from 20s data), taking the modality invariant representation  $Z$  as input. Our implementation uses two fully connected linear layers followed by a Sigmoid activation as the classification block.

## 4 EXPERIMENTS AND EVALUATION OF MI-MIL

This section evaluates the performance of our MI-MIL approach and different approach components. We compared MI-MIL's performance with state-of-the-art attention-based-MIL approach [35], DNN CNN, LSTM, and LSTM

Table 1. Evaluation of Q1: CWS vs. CWNS from scripted dataset: ‘window-1’ signal

(a) Evaluation using raw features						(b) Evaluation using change-score features					
Model	Accuracy	F1	precision	Recall	specificity	Model	Accuracy	F1	precision %	Recall	specificity
DNN	0.58	0.60	0.56	0.66	0.50	DNN	0.53	0.53	0.52	0.55	0.50
CNN	0.56	0.62	0.54	0.72	0.40	CNN	0.54	0.54	0.53	0.55	0.53
LSTM	0.69	0.65	0.74	0.59	0.80	LSTM	0.53	0.52	0.52	0.52	0.53
LSTM (with Attention)	0.69	0.65	0.74	0.59	0.80	LSTM (with Attention)	0.54	0.53	0.54	0.52	0.57
Attention-MIL	0.86	0.85	0.92	0.81	0.91	Attention-MIL	0.63	0.69	0.59	0.85	0.42
MI-MIL	0.88	0.90	0.83	0.98	0.79	MI-MIL	0.80	0.80	0.79	0.82	0.77

with attention approaches. The architectural information for MI-MIL, attention-based MIL and the mentioned baseline models are discussed in Appendix 8.5.1, 8.5.2 & 8.5.3 respectively. The presented network parameter configurations were optimized by performing a grid search of the possible network-parameters.

Following we present the dataset splits and evaluation metrics used for our evaluations. Later, section 4.1 discusses the models’ performance on the research question-wise tasks discussed in section 1.5 and the evaluation conclusions. Finally, section 4.2 evaluates the raw features’ importance in addressing the research questions and real-time executability of MI-MIL.

*Dataset Split and Evaluation Metrics:* For each of the evaluations, we followed *the person-disjoint hold-out method* [15]. We divided each dataset into three person-disjoint train, validation, and test sets, randomly selecting an equal number of participants from the CWS and CWNS groups. The distribution was as follows: test set (all data from 3-CWS and 3-CWNS), validation set (all data from 3-CWS and 3-CWNS), and training set (rest of the data). The same training, validation, and test set distributions were used for all evaluations of a dataset. To reduce contingency and avoid overfitting, classifiers were trained (on training+validation set) and evaluated (on the test set) three times with different seed values, and the average results are reported in this paper. Evaluation results are presented with the metrics: recall, precision, accuracy, F1-score, and specificity.

#### 4.1 Investigating the Research Questions through Classification Evaluation

We evaluated MI-MIL binary classifiers for each of the questions. We trained different binary classification models using two sets of input features: raw physiological features (section 2.3.1) and change-score features (section 2.3.2).

**4.1.1 Evaluation of Q1: Differentiating the CWS vs. CWNS from Scripted ‘window-1’ Signal:** To get insights for our research question 1, we evaluated the models to differentiate the CWS vs. CWNS from ‘window-1’ of the scripted dataset. ‘Window-1’ particularly comprises the participants’ physiological responses upon exposure to external stressors (*negatively-valenced picture*). Tables 1a & 1b show our evaluation results.

The MI-MIL approach achieves F1-scores of 0.90 and 0.80 (for raw and change score features respectively), establishing that CWS exhibit easily identifiable unique physiological ‘attributes’ (raw features) and ‘fluctuation-from-neural-condition’ (change-score features) patterns than the CWNS while exposed to external stressors.

Notably, MI-MIL outperforms all the baseline models. Specifically, in table 1b, the lower performance of the Attention-based-MIL indicates that the existing CWS distinctive sparse situational fluctuation-from-neural-condition patterns are subtle and more disjoint across different modalities compared to the physiological attribute patterns. However, MI-MIL’s modality-specific embedding blocks identify the patterns effectively, resulting in higher performance. Supervised learning baseline approaches perform relatively poorly since they fail to optimize with the absence of fine-grain data annotations, hence confirming the need for a weakly supervised learning MIL paradigm for effective classification.

Table 2. Evaluation of Q2: CWS vs. CWNS from scripted dataset: ‘window-2’ signal

(a) Evaluation using raw features						(b) Evaluation using change-score features					
Model	Accuracy	F1	precision	Recall	specificity	Model	Accuracy	F1	precision	Recall	specificity
DNN	0.63	0.67	0.59	0.76	0.50	DNN	0.58	0.55	0.58	0.52	0.63
CNN	0.61	0.69	0.57	0.90	0.33	CNN	0.58	0.56	0.57	0.55	0.60
LSTM	0.80	0.79	0.81	0.76	0.83	LSTM	0.58	0.55	0.58	0.52	0.63
LSTM (with Attention)	0.78	0.76	0.81	0.72	0.83	LSTM (with Attention)	0.64	0.60	0.67	0.55	0.73
Attention-MIL	0.88	0.84	0.91	0.80	0.91	Attention-MIL	0.72	0.76	0.66	0.90	0.56
MI-MIL	0.88	0.89	0.82	0.98	0.78	MI-MIL	0.83	0.83	0.82	0.85	0.80

Table 3. Evaluation of Q3: CWS vs. CWNS from Scripted baseline signal

Model	Accuracy	F1	precision	Recall	specificity
DNN	0.65	0.64	0.68	0.61	0.70
CNN	0.63	0.69	0.68	0.70	0.48
LSTM	0.63	0.67	0.61	0.75	0.50
Attention-MIL	0.58	0.65	0.65	0.73	0.43
MI-MIL	0.51	0.56	0.52	0.62	0.40

**4.1.2 Evaluation of Q2: Differentiating the CWS vs. CWNS from Scripted ‘window-2’ Signal:** To get insights for research question 2, we evaluated the models to differentiate the CWS vs. CWNS from ‘window-2’ of the scripted dataset. In ‘window-2’ of the scripted dataset, participants are repeating the predetermined BBAP phrase. Hence, this window comprises the participants’ physiological responses while talking under stressful conditions. Our evaluation results using raw and change-score features are shown in tables 2a & 2b.

The MI-MIL approach achieves F1-scores of 0.89 and 0.83 for raw and change score features respectively, establishing that CWS exhibits unique physiological attributes (i.e., amplitudes) and fluctuation-from-neutral-condition patterns than the CWNS while talking in stressful conditions. Comparing Table 1b & 2b, demonstrates that, the CWS’s unique physiological parameters fluctuation patterns are more explicit and identifiable while talking in stressful conditions. MI-MIL significantly outperforms supervised learning baselines (i.e., CNN, DNN, LSTM, LSTM with attention) and MIL baseline (attention-based-MIL), demonstrating that the presence of sparse and modality-specific disjoint physiological response patterns in CWS while they speak under stressful condition.

**4.1.3 Evaluation of Q3: Differentiating the CWS vs. CWNS from Scripted Dataset Baseline Signal:** This evaluation addresses the research question 3. It is important to understand if the CWS and CWNS show differences in physiology during their neutral affective state or baseline condition. If yes, the use of raw physiological features for classification may provide erroneous insights. For example, if CWS have higher physiological parameter values than the CWNS in baseline/neutral condition, identifying that the CWS and CWNS are showing similar parameter values (i.e., low classification accuracy) in a challenging situation would not demonstrate that they have similar physiological responses. Instead, it may indicate the CWNSs’ parameters fluctuation is higher (i.e., stronger physiological response) than the CWSs’.

Hence, we developed and evaluated models to differentiate CWS vs. CWNS from the 20s baseline signals of the scripted dataset. The evaluation results are shown in table 3. All models take raw physiological features as input and achieve low F1-scores (0.56-0.69). These results demonstrate that the CWS and CWNS exhibit similar physiological parameters during neutral/baseline conditions, though some subtle differentiating patterns exist.

Such findings justify our evaluation of the models using raw and change-score features separately. Models with raw features give us insights into the physiological parameters value differences. Models with change-score features give us insights into the fluctuations in physiology that represent stronger or weaker responses (i.e., higher or lower fluctuations).

Table 4. Evaluation of Q4: CWS vs. CWNS from Free-speech dataset

(a) Evaluation using raw features						(b) Evaluation using Change score features					
Model	Accuracy	F1	precision	Recall	specificity	Model	Accuracy	F1	precision	Recall	specificity
DNN	0.43	0.60	0.44	0.93	0.01	DNN	0.55	0.42	0.50	0.36	0.71
CNN	0.55	0.66	0.50	0.94	0.23	CNN	0.50	0.43	0.44	0.41	0.57
LSTM	0.56	0.63	0.51	0.83	0.33	LSTM	0.42	0.41	0.38	0.46	0.39
LSTM (with Attention)	0.56	0.57	0.51	0.64	0.50	LSTM (with Attention)	0.46	0.53	0.44	0.68	0.28
Attention-MIL	0.70	0.69	0.66	0.74	0.67	Attention-MIL	0.66	0.64	0.61	0.71	0.62
MI-MIL	0.72	0.73	0.68	0.85	0.62	MI-MIL	0.64	0.67	0.57	0.81	0.49

**4.1.4 Evaluation of Q4: Differentiating the CWS vs. CWNS from Free-Speech Signal:** Though no external stressors were imposed during the free-speech experiment, the narration task is linguistically and cognitively demanding. It may elicit different physiological responses in CWS vs. CWNS. To investigate such differences in responses, we evaluated the models to detect the differences in the CWS vs. CWNS from the free-speech dataset. Our evaluation results using raw and change-score features are shown in tables 4a & 4b.

The MI-MIL approach achieves F1-scores of 0.73 and 0.67 for raw and change score features respectively, establishing that CWS exhibit identifiable unique physiological responses than the CWNS while performing linguistically and cognitively demanding tasks (i.e., narration). MI-MIL approach outperforms all the baseline models, demonstrating its higher efficacy in identifying subtle, sparse and modality-disjoint patterns.

Notably, no predetermined phrases or sequences of phonemes were uttered during the spontaneous narration task. Hence, differentiating CWS vs. CWNS was more challenging than the scripted dataset's 'BBAP' phrase repetition task. Therefore, this section's lower evaluation accuracy (table 4) compared to the tables 1 & 2, does not indicate that the CWS exhibit a more explicit or stronger physiological response difference than CWNS in the stressful talking task compared to the narration task.

## 4.2 Investigating the Approach Components

This section evaluates the raw features' importance and real-time executability of MI-MIL.

**4.2.1 Feature Selection:** This section evaluates the discriminative capability of the raw physiological features for the CWNS vs. CWS classification task. Notably, recent literature [9, 58, 66, 71, 74] have established that the HR, EDA, RSP-rate, RSP-amp are effective features for affective states and stuttering individual's physiological arousal detection. Hence, this section aims to identify the features' relative importance rather than feature selection. Following prior studies [22, 29, 82, 93, 94], we employ ridge-regression. Ridge regression adds 'squared magnitude' of coefficient as penalty term to the loss function, hence highly penalizes coefficient of less important features. Evaluations are done on the scripted dataset for a fair comparison, where all participants experienced a similar condition. Table 5a presents the feature rankings. EDA features are highly important, following RSP-rate, HR, and RSP-amp features. These results are in line with our ML interpretability evaluations.

We also evaluate the performance of our baseline Attention MIL model utilizing the top-k features from the ranks, and the results are presented in Table 5b. Notably, since not many modalities were selected during Top-5, -10 evaluations, MI-MIL was not utilized. According to our evaluation, adding more features result in higher performance. These evaluations indicate that our classifiers were not overfitting due to redundant features, which is obvious since we have only 24 features, and all of them are shown to be effective by the literature.

**4.2.2 Execution Time and Resource Usage on Scalable Devices:** We evaluated MI-MIL's real-time executability and resource usage on scalable mobile devices: an Nvidia Jetson Nano and a smartphone. We run the MI-MIL models taking consecutive 20-sec windows for 10 minutes and record the running time and resource usage. According to the Tablse 6, MI-MIL takes 0.019 sec and 0.005 sec on Jetson Nano and smartphone to process each 20-sec

Table 5. Feature rankings based on coefficients of ridge classifier and classification performance utilizing top- $k$  features.

(a) Feature selection rankings

Rank	Feature name	Coefficient	Rank	Feature name	Coefficient
1	EDA Min	0.468	13	HR Median	-0.002
2	EDA Max	0.191	14	HR Min	-0.012
3	EDA Max	0.159	15	HR Max	-0.017
4	RSP rate Mean	0.151	16	HR Variance	-0.021
5	EDA Mean	0.057	17	RSP amp Median	-0.023
6	EDA Variance	0.056	18	RSP amp STD	-0.045
7	RSP rate Max	0.038	19	RSP rate STD	-0.048
8	HR STD	0.031	20	RSP amp Mean	-0.065
9	RSP amp Variance	0.065	21	RSP rate Min	-0.066
10	RSP amp Min	0.012	22	EDA STD	-0.079
11	HR Mean	0.008	23	RSP rate Median	-0.089
12	RSP rate Variance	0.003	24	EDA Median	-0.202

(b) Performance of Attention MIL for Top- $k$  features  
( $k=5, 10, 15, 20$ )

Model	Accuracy	F1	precision	Recall	specificity
Attention-MIL (Top-5)	0.77	0.79	0.72	0.89	0.66
Attention-MIL (Top-10)	0.83	0.84	0.77	0.91	0.75
Attention-MIL (Top-15)	0.84	0.82	0.95	0.72	0.96
Attention-MIL (Top-20)	0.85	0.86	0.78	0.96	0.75

physiological data window. The average CPU and GPU usages are also low. The results suggest that MI-MIL can perform real-time analysis on resource constraint devices.

Table 6. Evaluation of MI-MIL on Jetson Nano and Google pixel 6. The Jetson nano is equipped with NVIDIA Maxwell GPU, Quad-core ARM processor, and 4GB memory. The smartphone is a google pixel 6, powered by Octa-core ( 2x2.80 GHz Cortex-X1 , 2x2.25 GHz Cortex-A76 and 4x1.80 GHz Cortex-A55) with 8GB memory.

(a) Evaluation of MI-MIL on Jetson Nano

Features	Scripted raw	Scripted change-score	Free speech raw	Free speech change-score
Average CPU usage (%)	10.90	11.13	10.82	10.3
Average GPU usage (%)	0.07	0.108	0.124	0.09
Average memory usage (Mb)	1703	1692	1688	1538
Run-time (s)	0.017	0.018	0.018	0.021

(b) Evaluation of MI-MIL on Google pixel 6

Features	Scripted raw	Scripted change-score	Free speech raw	Free speech change-score
Average CPU usage (%)	2	4	6	6
Average memory usage (Mb)	166	169	167	128
Run-time (s)	0.0061	0.0043	0.0058	0.0055

## 5 INTERPRETABILITY VISUALIZATION AND DISCUSSION

MI-MIL's inferences can be utilized to understand stuttering children's psychological arousal during speaking from a *group-wise* and *personalized* perspective. The developed binary ML classifiers in section 4.1 are BlackBox. Understanding which physiological features (raw or change-scores) are important for the respective model's inference is critical since they contribute most to differentiating CWSs' physiological response from others. We employed the KernelSHAP, a model-agnostic interpretation framework that determines each physiological feature's (raw or change-score) impact in terms of its Shapley value [67] on the respective model's inferences. Notably, distinct Shapley feature importance values are generated for each input, indicating each feature's impact on generating a class inference for that particular input (*i.e., personalized perspective*).

However, Shapley values are additive [51]. We average the Shapley values per feature across the data to consider the *global importance*. Global feature importance indicates how much a model relies on each feature at each timestamp, overall (*i.e., group-wise perspective*). We calculate the global importance for a model's true predictions by computing the mean of the generated Shapley values corresponding to the test set 20s windows.

We present the global feature importance for a research-question respective MI-MIL model's inference through a grid-heatmap (with a cell for each features in each timestamp, e.g., figure 5). In contrast, we present the personalized feature importance of each 20s physiological response data (from a CWS or CWNS) through time-series representation of the features and heatmap on each 2s segment of the time-series (section 5.1.2).

The heatmaps use blue color to show positive Shapley values (responsible for pushing the model towards CWNS), and red color for negative Shapley values (responsible for pushing the model decision toward CWS),

darker the intensity of the red/blue color higher the magnitude of the Shapley value and higher is the importance of the feature in pushing the model towards CWS/CWNS class.

Speech science studies' interest lies in understanding the second-to-second effect of children's physiological arousal in their speech production; hence we are particularly interested in visualizing and understanding the importance of physiological features in differentiating CWS vs. CWNS while repeating the target phrase (BBAP) after viewing negatively-valenced (stress-provoking) pictures (discussed in section 5.1), and while describing pictures spontaneously during the free-speech condition, discussed in Appendix 8.4.

### 5.1 Q2 Interpretation: MI-MIL Model to Differentiate CWS vs. CWNS While Speaking in Stress Condition (Scripted Dataset)

This section discusses and demonstrates how MI-MIL's inferences can be utilized and visualized to understand stuttering children's distinctive psychological responses during speaking under stress condition from a group-wise (section 5.1.1) and personalized (section 5.1.2) level. The ML interpretations are discussed in following:

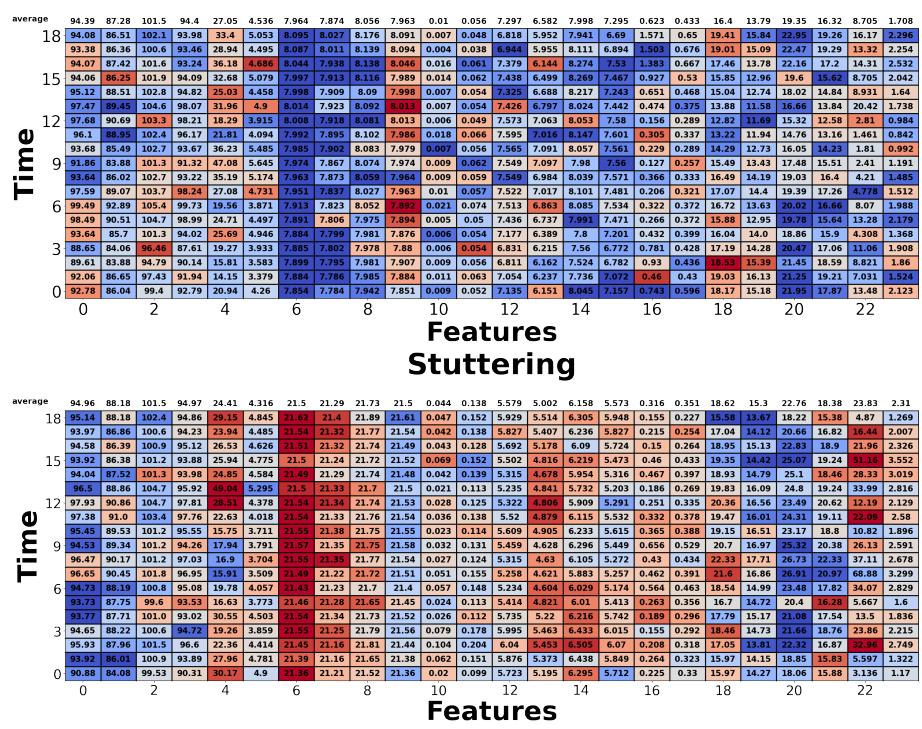
**5.1.1 Group-wise Global Feature Importance:** The figures 5a and 5b show the global features importance through Shapley interpretation graphs of the MI-MIL models (section 4.1.2) evaluated on window-2. Each of the figures comprises two SHAP plots: features' importance plot for true negative (i.e., detecting CWS) and true positive classification (i.e., detecting CWNS). Window-2 comprises children's physiology during the BBAP phrase repetition under stressful conditions task, and the evaluation in section 4.1.2 indicates that the CWS and CWNS show significant differences. This section visualizes and discusses the MI-MIL identified group-wise differences in CWS vs. CWNS through the respective model's interpretation.

**EDA Features Importance:** As shown in figure 5a, CWS experienced higher raw EDA feature values than CWNS, indicating CWS experienced higher physiological arousal [9, 58]. These raw EDA features are important in CWS vs. CWNS classification, indicating that they are distinctive patterns. Psychophysiological research examining speaking-related sympathetic nervous system activity in CWS is limited to just a few studies. However, speaking task- and age-related differences in skin conductance level have been reported [18, 36, 37, 72, 74, 89, 90]. *This study is the first to evident that, the CWS experience distinctively higher arousal while talking in stressful condition than their non-stuttering peers.* When interpreting our raw EDA findings, the readers are reminded that in this evaluation, we used raw EDA data (phasic skin conductance responses were not removed from the signal; similarly, the baseline EDA was not considered in this model).

**HR Features Importance:** According to the figure 5a, for CWS participants, the HR variance during the later part of the 20s window (timesteps 12-18) shows a sudden increase and has high importance toward CWS classification (indicated by the dark red cells). Research indicates that social-emotional challenges or a feeling of anxiety elicit increased heart rate attributes in children and adults [45]. Hence, *we interpret the discussed HR-variance pattern as a sign that during talking under stressful conditions, CWS's arousal increases with time, and it is a distinctive attribute between CWS and CWNS.* Such interpretation is in line with the findings from the stuttering research literature. A recent statistical analysis-based study [74] showed that CWS experienced on avg. higher HR attributes, hence, higher emotional arousal than CWNS. *In contrast, our approach allows finding second by second temporal patterns in specific modalities like HR, which are distinctive in CWS.*

**RSP Features Importance:** In regard to the raw respiratory effort data (per the figure 5b), RSP-rate mean and variance contributed significantly to differentiating CWS vs. CWNS. Although the raw respiratory effort values were higher in CWS than in CWNS, the data for both groups are in line with speech breathing rates for preschool-age children [12]. Thus, *our finding indicates the relatively faster speech breathing rates of CWS (compared to CWNS) as a distinctive respiratory effort pattern.* Interestingly, CWSs' RSP feature values slightly decrease with

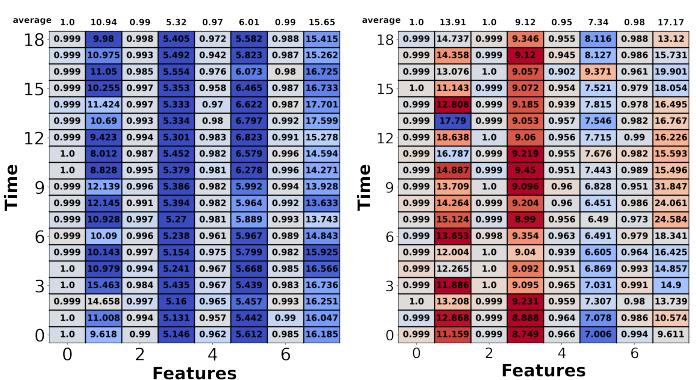
### Non-Stuttering



(a) Window-2 Raw features

### Non-Stuttering

### Stuttering



(b) Window-2 change-score features

Fig. 5. Shapley global features-importance heatmaps for MI-MIL models evaluated on (Window-2) scripted dataset. In each plot, the X-axis represents the features (24 raw physiological and 8 change score features), and the Y-axis represents the '2s segments' inside the 20s event detection window. Each cell in the heatmap contains the mean Shapley value of the respective feature in that 2s segment. The color-bars in the right of the figures show the colors scheme according to Shapley value's magnitude. The raw and change-score features, with their corresponding indexes on the generated Shapley feature importance plots are shown in Appendix 8.2 - table 8.

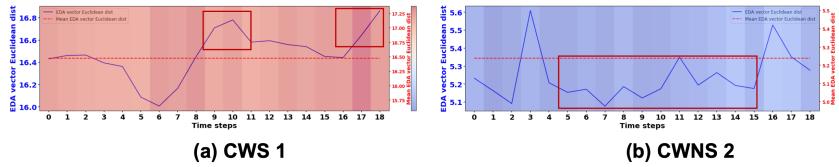


Fig. 6. Participant wise shapley importance plots for EDA change score euclidean distance features (from one window-2)

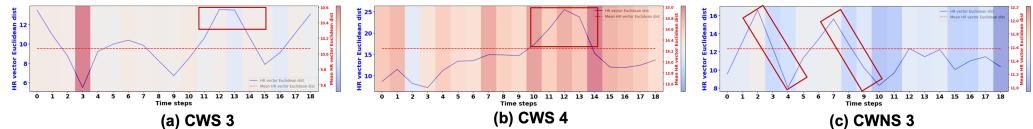


Fig. 7. Participant wise shapley importance plots for HR change score euclidean distance features (from one window-2)

time progression (in Y-axis), *indicating that with the progression of time, CWS's speech breathing rate decreases even under stressful conditions*. Though, it has been proposed that stuttering is associated with various airflow irregularities (e.g., [11]), but to our knowledge, this is the first study to examine respiratory effort in preschool-age children who stutter.

**Change-score Features Importance:** We also evaluated the change-score of the four physiological features (HR, EDA, RSP-amp, and RSP-rate). According to the figures 5b, CWS showed a higher fluctuation from baseline in HR, EDA, RSP-amplitude, and RSP-rate features compared to CWNS. HR and EDA change score features have higher shapley importance. Hence, CWS showing higher fluctuation in EDA and HR during talking under stressful condition compared to the CWNS are distinctive patterns. It demonstrates that CWS experience higher arousal than their peers while speaking in stressful conditions.

According to this section's evaluation and interpretation, we can conclude that the CWS showed distinctive temporal and overall-window-wide physiological parameter differences from their non-stuttering peers. This paper's presented approach can identify these patterns in fine-grain, second by second level.

**5.1.2 Personalized Interpretation:** On average, across CWS participants, the HR and EDA changes-scores from window-2 are distinctive. A major contribution of this paper is its ability to extract and visualize personalized fine-grain second by second temporal physiological response patterns. Figures 6 and 7 demonstrate the visualization of personalized EDA and HR feature importance in 20s-window-2 from five different individuals. According to figure 6, 'CWS 1' participant has higher EDA, meaning experiencing higher arousal [9, 58] than the 'CWNS 2' participant. Moreover, the 'CWS 1' participant's arousal increases with time, indicating talking in stressful situations in enhancing this participant's arousal more and more. Notably, the MI-MIL Shapley feature importance values are higher (darker 'red') on those increasing EDA picks, indicating the MI-MIL approach can effectively identify the personalized patterns.

According to figure 7, both CWS participants' HR change-score is increasing with time, which is in line with section 5.1.1. For 'CWS 4', the HR-change-score pick is a high value (timestamp 10-14) and has high Shapley feature importance values, similar to the CWS group. However, 'CWS 3' individual's response is different. 'CWS 3' experienced a 'freezing response' at the beginning of talking (timestamp 3); hence the overall HR change-score values are not as high as the other CWS. Notably, the MI-MIL Shapley feature importance values are higher (darker 'red') on timestamp 3 (during the freezing response), indicating the MI-MIL approach can effectively identify the distinctive personalized patterns.

Additionally, in figure 7(c), the 'CWNS 3' individual's HR change-score is relatively lower, has decreasing trends where the Shapley importance values are higher, which indicates the distinctive pattern of 'CWNS 3' individual that indicates the participant is not from the stuttering group.

*This section's discussion demonstrates that our approach can effectively identify fine-grain, personalized, distinctive temporal patterns from CWS and CWNS individuals.* Identifying such patterns would enable personalized understanding of stuttering development and potential just-in-time personalized interventions to mitigate the physiological responses that may affect a children's speech.

## 6 STUDY OBSERVATIONS, IMPACT, AND LIMITATIONS

We developed a novel MI-MIL approach (section 3.2) that addresses the challenges present in differentiating CWS vs. CWNS's situational physiological arousal (section 1.4) from 'weakly labeled' data. MI-MIL's high efficacy in addressing all research questions indicates its effectiveness.

Our evaluation visualization in section 5.1 demonstrates presented papers approach can effectively identify fine-grain, personalized, distinctive temporal patterns from CWS and CWNS, group-wise and personalized for each 20s window. Notably, many of the patterns in CWS's physiological response patterns are investigated for the first time. Many of them conform to the existing speech science literature showing the reliability of our approach's visualization. The following discusses the presented approach's impact, generalizability, and limitations.

*Impact and Application of the Study:* Approximately 5 percent of all children go through a period of stuttering, and 1 percent suffer from long-term stuttering [2]. Speech production is a complex process that requires fast and precise coordination of respiration, voice production, lip, tongue, and jaw movements (among other speech articulators) while simultaneously processing cognitive-linguistic information. Social engagement, including regulating own emotions and responding appropriately to one's communicative partner, is also inherent to spoken communication. Naturally, speech production can be affected by the speaker's physiological arousal. Young children who have speech disorders, such as stuttering, are especially vulnerable to these influences.

The fast nature of speech production calls for fine-grain, second-by-second assessment of any physiological response parameter of interest that can influence speech characteristics. To attain this goal, the presented study offers a new way to examine CWS's physiological arousal data and has both *group-level* and *personalized-level* impacts on stuttering and individuals with stuttering.

*Group-wise Impact:* This study's presented ML-based group-wise examination (discussed in section 5.1.1) of attributes and fluctuations in physiological arousal during speaking could inform our understanding of the role of physiological arousal in the development of stuttering and explain the origins of its situational variability, one of the key and unexplained features of this condition.

*Personalized-level Impact:* Recent literature [6, 23, 39, 43] suggests that CWS (compared with fluent peers) have increased difficulty in the regulation and/or adaptation of their behavioral and emotional responses to everyday scenarios which lead to increased emotional reactivity to stressful stimuli. This paper's evaluation not only confirms that, moreover, the presented approach can identify the fine-grain, second by second, temporal and personalized distinctive physiological response patterns of each CWS.

Important to note that the physiological sensing parameters evaluated in this study are present in recent wearables. For example, same Biopac sensors utilized in this study (section 2) are present in Biopac wireless wearable system [1]; hence the MI-MIL models can be implemented and evaluated in wearables. According to our evaluation in section 4.2.2, developed MI-MIL models are real-time executable in smart devices (e.g., smartphones). Hence, the presented approach has the potential to be leveraged for remote, continuous, automated, and real-time assessment of stuttering children's physiological responses. Such assessments can be used clinically to facilitate providing just-in-time emotion regulation strategies (e.g., biofeedback, mindfulness interventions [59]) that may lead to improvement in speech fluency/disfluency.

*Generalizability of MI-MIL:* It is important to note that the MI-MIL approach is not limited to CWS's physiological response. To demonstrate the generalizability, we evaluated the MI-MIL model on a benchmark dataset

named WESAD dataset [65] which contains the physiological response parameters (RSP, ECG, EDA) from 15 participants for baseline and stress conditions. Notably, WESAD dataset is not weakly labeled; has fine-grain annotations. Still, the MI-MIL approach outperformed all the baselines: DNN, CNN, LSTM, and Attention-based MIL models by achieving an F1 score of 0.92 (Detail in Appendix 9), which shows its generalizability, robustness, and applicability in different domains.

*Study Limitations:* The study's limitation is that we analyzed data from only two conditions. Future work would benefit from sampling data from a broader range of situations to determine the models' predictive validity boundaries. Additionally, in the future, our approach can be implemented and evaluated on wearables for the longitudinal understanding of CWS's situational physiological arousal. Notably, this study's scope does not include an evaluation of CWS's speech. Multi-modal analysis of speech acoustics and physiological parameters can be a future research direction.

## 7 CONCLUSION

The presented first-of-its-kind study effectively identifies and visualizes the second-by-second temporal pattern differences in the physiological arousal of preschool-age CWS and CWNS while speaking perceptually fluently in two challenging conditions: speaking in stressful situations and narration. We collected physiological parameters data from 70 children in the two target conditions. However, our dataset and differentiating CWS from CWNS leveraging their physiological response has several challenges (section 1.4), which we address by developing a novel MI-MIL. MI-MIL applies a multiple-instance-learning paradigm on each modality independently, while through a cross-modality-fusion network, it effectively combines each modality's sparse, latent attributes. MI-MIL is real-time executable and generalizable to other domains. The evaluation of this classifier addresses four critical research questions that align with state-of-the-art speech science studies' interests. Later, we leverage SHAP classifier interpretations to visualize the salient and fine-grain physiological parameters unique to CWS. Finally, comprehensive evaluations are done on multiple datasets, presented framework, and several baselines that identified notable insights on CWSs' physiological arousal during speech production.

## ACKNOWLEDGMENTS

This work was supported in part by NSF IIS SCH # 2124285, NIH NIDCD # R21DC018103, and NIH NIDCD # R01DC017476 – S2. We extend our sincere gratitude to the children and their caregivers whose participation made this project possible.

## REFERENCES

- [1] [n.d.]. Biopac Wireless Wearable Physiology Sensing system. <https://www.biopac.com/product-category/research/bionomadix-wireless-physiology/>. Accessed: 2022-05-01.
- [2] [n.d.]. Stuttering population in USA. <https://www.stutteringhelp.org/prevalence>. Accessed: 2022-05-01.
- [3] Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2016. Human acute stress detection via integration of physiological signals and thermal imaging. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–8.
- [4] Fabrizio Albertetti, Alena Simalastar, and Aücha Rizzotti-Kaddouri. 2021. Stress Detection with Deep Learning Approaches Using Physiological Signals. In *IoT Technologies for HealthCare: 7th EAI International Conference, HealthyIoT 2020, Viana do Castelo, Portugal, December 3, 2020, Proceedings* 7. Springer International Publishing, 95–111.
- [5] Per A Alm. 2004. Stuttering, emotions, and heart rate during anticipatory anxiety: A critical review. *Journal of fluency disorders* 29, 2 (2004), 123–133.
- [6] Per A Alm. 2014. Stuttering in relation to anxiety, temperament, and personality: Review and analysis with focus on causality. *Journal of fluency disorders* 40 (2014), 5–21.
- [7] Hayley S Arnold, Megan K MacPherson, and Anne Smith. 2014. Autonomic correlates of speech versus nonspeech tasks in children and adults. *Journal of Speech, Language, and Hearing Research* 57, 4 (2014), 1296–1307.

- [8] Ali Azarbarzin, Michele Ostrowski, Patrick Hanly, and Magdy Younes. 2014. Relationship between arousal intensity and heart rate response to arousal. *Sleep* 37, 4 (2014), 645–653.
- [9] Dominik R Bach, Jean Daunizeau, Nadine Kuelzow, Karl J Friston, and Raymond J Dolan. 2011. Dynamic causal modeling of spontaneous fluctuations in skin conductance. *Psychophysiology* 48, 2 (2011), 252–257.
- [10] Stefanie Blain, Sarah D Power, Ervin Sejdic, Alex Mihailidis, and Tom Chau. 2010. A cardiorespiratory classifier of voluntary and involuntary electrodermal activity. *Biomedical engineering online* 9, 1 (2010), 1–18.
- [11] Oliver Bloodstein, Nan Bernstein Ratner, and Shelley B Brundage. 2021. *A handbook on stuttering*. Plural Publishing.
- [12] Carol A Boliek, Thomas J Hixon, Peter J Watson, and Patricia B Jones. 2009. Refinement of speech breathing in healthy 4-to 6-year-old children. (2009).
- [13] Margaret M Bradley and Peter J Lang. 2007. Emotion and motivation. (2007).
- [14] Marc-André Carboneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77 (2018), 329–353.
- [15] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11 (2010), 2079–2107.
- [16] Hugh Chen, Scott Lundberg, and Su-In Lee. 2019. Explaining Models by Propagating Shapley Values of Local Components. *arXiv preprint arXiv:1911.11888* (2019).
- [17] Hung-Chih Chiu, Yen-Hung Lin, Men-Tzung Lo, Sung-Chun Tang, Tzung-Dau Wang, Hung-Chun Lu, Yi-Lwun Ho, Hsi-Pin Ma, and Chung-Kang Peng. 2015. Complexity of cardiac signals for predicting changes in alpha-waves after stress in patients undergoing cardiac catheterization. *Scientific reports* 5, 1 (2015), 1–14.
- [18] Dahye Choi, Edward G Conture, Tedra A Walden, Robin M Jones, and Hanjoe Kim. 2016. Emotional diathesis, emotional stress, and childhood stuttering. *Journal of Speech, Language, and Hearing Research* 59, 4 (2016), 616–630.
- [19] Lei Clifton and David A Clifton. 2019. The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials* 20, 1 (2019), 1–6.
- [20] Alberto de Santos Sierra, Carmen Sánchez Ávila, Javier Guerra Casanova, and Gonzalo Bailador del Pozo. 2011. A stress-detection system based on physiological signals and fuzzy logic. *IEEE Transactions on Industrial Electronics* 58, 10 (2011), 4857–4865.
- [21] Alberto de Santos Sierra, Carmen Sanchez Avila, Javier Guerra Casanova, Gonzalo Bailador del Pozo, and Vicente Jara Vera. 2010. Two stress detection schemes based on physiological signals for real-time applications. In *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 364–367.
- [22] Natarajan Deepa, B Prabadevi, Praveen Kumar Maddikunta, Thippa Reddy Gadekallu, Thar Baker, M Ajmal Khan, and Usman Tariq. 2021. An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. *The Journal of Supercomputing* 77, 2 (2021), 1998–2017.
- [23] Kurt Eggers, F Luc, and Bea RH Van den Bergh. 2013. Inhibitory control in childhood stuttering. *Journal of fluency disorders* 38, 1 (2013), 1–13.
- [24] James Foulds and Eibe Frank. 2010. A review of multi-instance learning assumptions. *The knowledge engineering review* 25, 1 (2010), 1–25.
- [25] Eric Golinko and Xingquan Zhu. 2019. Generalized feature embedding for supervised, unsupervised, and online learning tasks. *Information Systems Frontiers* 21, 1 (2019), 125–142.
- [26] Raul Gomez, Lluis Gomez, Jaume Gibert, and Dimosthenis Karatzas. 2019. Self-supervised learning from web data for multimodal retrieval. In *Multimodal Scene Understanding*. Elsevier, 279–306.
- [27] C Goumopoulos and E Menti. 2019. Stress Detection in Seniors Using Biosensors and Psychometric Tests. *Procedia Computer Science* 152 (2019), 18–27.
- [28] Tsafir Greenberg, Joshua M Carlson, Denis Rubin, Jook Cha, and Lilianne Mujica-Parodi. 2015. Anticipation of high arousal aversive and positive movie clips engages common and distinct neural substrates. *Social cognitive and affective neuroscience* 10, 4 (2015), 605–611.
- [29] WANG Guodong, SUN Lanxiang, WANG Wei, CHEN Tong, GUO Meiting, and Peng Zhang. 2020. A feature selection method combined with ridge regression and recursive feature elimination in quantitative analysis of laser induced breakdown spectroscopy. *Plasma Science and Technology* 22, 7 (2020), 074002.
- [30] SL Happy, Antitza Dantcheva, and Francois Bremond. 2019. A Weakly Supervised learning technique for classifying facial expressions. *Pattern Recognition Letters* 128 (2019), 162–168.
- [31] Cindy Harmon-Jones, Brandon J Schmeichel, Eileen Mennitt, and Eddie Harmon-Jones. 2011. The expression of determination: similarities between anger and approach-related positive affect. *Journal of personality and social psychology* 100, 1 (2011), 172.
- [32] Heather A Henderson, Daniel S Pine, and Nathan A Fox. 2015. Behavioral inhibition and developmental risk: a dual-processing perspective. *Neuropsychopharmacology* 40, 1 (2015), 207–224.
- [33] S Het, N Rohleder, D Schoofs, C Kirschbaum, and OT Wolf. 2009. Neuroendocrine and psychometric evaluation of a placebo version of the ‘Trier Social Stress Test’. *Psychoneuroendocrinology* 34, 7 (2009), 1075–1086.

- [34] Murtadha D Hssayeni and Behnaz Ghoraani. 2021. Multi-Modal Physiological Data Fusion for Affect Estimation Using Deep Learning. *IEEE Access* 9 (2021), 21642–21652.
- [35] Maximilian Ilse, Jakub M Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712* (2018).
- [36] Robin M Jones, Anthony P Buhr, Edward G Conture, Victoria Tumanova, Tedra A Walden, and Stephen W Porges. 2014. Autonomic nervous system activity of preschool-age children who stutter. *Journal of fluency disorders* 41 (2014), 12–31.
- [37] Robin M Jones, Tedra A Walden, Edward G Conture, Aysu Erdemir, Warren E Lambert, and Stephen W Porges. 2017. Executive functions impact the relation between respiratory sinus arrhythmia and frequency of stuttering in young children who do and do not stutter. *Journal of Speech, Language, and Hearing Research* 60, 8 (2017), 2133–2150.
- [38] James D Keeler, David E Rumelhart, and Wee Kheng Leow. 1991. Integrated segmentation and recognition of hand-printed numerals. In *Advances in neural information processing systems*. 557–563.
- [39] Elaina Kefalianos, Mark Onslow, Susan Block, Ross Menzies, and Sheena Reilly. 2012. Early stuttering, temperament and anxiety: Two hypotheses. *Journal of fluency disorders* 37, 3 (2012), 151–163.
- [40] Joni Kettunen, Niklas Ravaja, Petri Näätänen, Pertti Keskivaara, and Liisa Keltikangas-Järvinen. 1998. The synchronization of electrodermal activity and heart rate and its relationship to energetic arousal: A time series approach. *Biological Psychology* 48, 3 (1998), 209–225.
- [41] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The ‘Trier Social Stress Test’—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.
- [42] Jennifer Kleinow and Anne Smith. 2000. Influences of length and syntactic complexity on the speech motor stability of the fluent speech of adults who stutter. *Journal of Speech, Language, and Hearing Research* 43, 2 (2000), 548–559.
- [43] Shelly Jo Kraft, Emily Lowther, and Janet Beilby. 2019. The role of effortful control in stuttering severity in children: Replication study. *American Journal of Speech-Language Pathology* 28, 1 (2019), 14–28.
- [44] Marzena Kryszkiewicz. 2014. The cosine similarity in terms of the euclidean distance. In *Encyclopedia of Business Analytics and Optimization*. IGI Global, 2498–2508.
- [45] BM Kudielka, A Buske-Kirschbaum, DH Hellhammer, and C Kirschbaum. 2004. HPA axis responses to laboratory psychosocial stress in healthy elderly adults, younger adults, and children: impact of age and gender. *Psychoneuroendocrinology* 29, 1 (2004), 83–98.
- [46] Christof Kuhbandner and Michael Zehetleitner. 2011. Dissociable effects of valence and arousal in adaptive executive control. *PloS one* 6, 12 (2011), e29287.
- [47] Peter J Lang. 1995. The emotion probe: Studies of motivation and attention. *American psychologist* 50, 5 (1995), 372.
- [48] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1, 39–58 (1997), 3.
- [49] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (1993), 261–273.
- [50] Bing Liu. 2017. Many facets of sentiment analysis. In *A practical guide to sentiment analysis*. Springer, 11–39.
- [51] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*. 4765–4774.
- [52] Valentina Magagnin, Maurizio Mauri, Pietro Cipresso, Luca Mainardi, Emery N Brown, Sergio Cerutti, Marco Villamira, and Riccardo Barbieri. 2010. Heart rate variability and respiratory sinus arrhythmia assessment of affective states by bivariate autoregressive spectral analysis. In *2010 Computing in Cardiology*. IEEE, 145–148.
- [53] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. 2019. DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks.. In *AAMAS*. 2108–2110.
- [54] Maurizio Mauri, Valentina Magagnin, Pietro Cipresso, Luca Mainardi, Emery N Brown, Sergio Cerutti, Marco Villamira, and Riccardo Barbieri. 2010. Psychophysiological signals associated with affective states. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 3563–3566.
- [55] Mercer Mayer. 1974. Frog goes to dinner. 1 (1974).
- [56] F. Mokhayeri, M. Akbarzadeh-T, and S. Toosizadeh. 2011. Mental stress detection using physiological signals based on soft computing techniques. In *2011 18th Iranian Conference of Biomedical Engineering (ICBME)*. 232–237.
- [57] Oscar Martinez Mozos, Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, Radu Dobrescu, and Jose Manuel Fernandez. 2017. Stress detection using wearable physiological and sociometric sensors. *International journal of neural systems* 27, 02 (2017), 1650041.
- [58] Jelena Nikolic-Popovic and Rafik Goubran. 2011. Measuring heart rate, breathing rate and skin conductance during exercise. In *2011 IEEE International Symposium on Medical Measurements and Applications*. IEEE, 507–511.
- [59] Laura W Plexico and Mary J Sandage. 2011. A mindful approach to stuttering intervention. *Perspectives on Fluency and Fluency Disorders* 21, 2 (2011), 43–49.

- [60] Jan Ramon and Luc De Raedt. 2000. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*. 53–60.
- [61] Mohammad Naim Rastgoo, Bahareh Nakisa, Frederic Maire, Andry Rakotonirainy, and Vinod Chandran. 2019. Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications* 138 (2019), 112793.
- [62] A Rehman, M Mustafa, and I Israr. 2013. Survey of wearable sensors with comparative study of noise reduction ecg filters. *International Journal of Computing and Network Technology* 221, 1249 (2013), 1–21.
- [63] Asif Salekin, Zeya Chen, Mohsin Y Ahmed, John Lach, Donna Metz, Kayla De La Haye, Brooke Bell, and John A Stankovic. 2017. Distant emotion recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–25.
- [64] Asif Salekin, Jeremy W Eberle, Jeffrey J Glenn, Bethany A Teachman, and John A Stankovic. 2018. A weakly supervised learning framework for detecting social anxiety and depression. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 2 (2018), 1–26.
- [65] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–408.
- [66] Rainer Schneider, Stefan Schmidt, Markus Binder, Florian Schäfer, and Harald Walach. 2003. Respiration-related artifacts in EDA recordings: introducing a standardized method to overcome multiple interpretations. *Psychological reports* 93, 3 (2003), 907–920.
- [67] Lloyd S Shapley. 1953. Stochastic games. *Proceedings of the national academy of sciences* 39, 10 (1953), 1095–1100.
- [68] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.
- [69] Pekka Siirtola. 2019. Continuous stress detection using the sensors of commercial smartwatch. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1198–1201.
- [70] Anne Smith and Christine Weber. 2017. How stuttering develops: The multifactorial dynamic pathways theory. *Journal of Speech, Language, and Hearing Research* 60, 9 (2017), 2483–2505.
- [71] Stephen M Tasko, Michael D McClean, and Charles M Runyan. 2007. Speech motor correlates of treatment-related changes in stuttering severity and speech naturalness. *Journal of Communication Disorders* 40, 1 (2007), 42–65.
- [72] Victoria Tumanova and Nicole Backes. 2019. Autonomic nervous system response to speech production in stuttering and normally fluent preschool-age children. *Journal of Speech, Language, and Hearing Research* 62, 11 (2019), 4030–4044.
- [73] Victoria Tumanova, Edward G Conture, E Warren Lambert, and Tedra A Walden. 2014. Speech disfluencies of preschool-age children who do and do not stutter. *Journal of communication disorders* 49 (2014), 25–41.
- [74] Victoria Tumanova, Blair Wilder, Julia Gregoire, Michaela Baratta, and Rachel Razza. 2020. Emotional Reactivity and Regulation in Preschool-Age Children Who Do and Do Not Stutter: Evidence From Autonomic Nervous System Measures. *Frontiers in Human Neuroscience* 14 (2020).
- [75] Peter H Venables and Margaret J Christie. 1980. Electrodermal activity. *Techniques in psychophysiology* 54, 3 (1980).
- [76] MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal* 3, 2 (2016), 19–28.
- [77] Bruno Villegas, Kevin M Flores, Kevin José Acuña, Kevin Pacheco-Barrios, and Dante Elias. 2019. A novel stuttering disfluency classification system based on respiratory biosignals. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4660–4663.
- [78] Bridget Walsh, Anne Smith, Sharon L Christ, and Christine Weber. 2019. Sympathetic nervous system activity in preschoolers who stutter. *Frontiers in human neuroscience* 13 (2019), 356.
- [79] Bridget Walsh and Evan Usler. 2019. Physiological correlates of fluent and stuttered speech production in preschool children who stutter. *Journal of Speech, Language, and Hearing Research* 62, 12 (2019), 4309–4323.
- [80] Liwei Wang, Yan Zhang, and Jufu Feng. 2005. On the Euclidean distance of images. *IEEE transactions on pattern analysis and machine intelligence* 27, 8 (2005), 1334–1339.
- [81] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74 (2018), 15–24.
- [82] Zhi-Ming WANG, Na HAN, Zhe-Ming YUAN, and Zhao-Hua WU. 2013. Feature selection for high-dimensional data based on ridge regression and SVM and its application in peptide QSAR modeling. *Acta Physico-Chimica Sinica* 29, 3 (2013), 498–507.
- [83] Christine M Weber and Anne Smith. 1990. Autonomic correlates of stuttering and speech assessed in a range of experimental tasks. *Journal of Speech, Language, and Hearing Research* 33, 4 (1990), 690–706.
- [84] Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Hermie Hermens, and Julien Penders. 2011. Towards mental stress detection using wearable physiological sensors. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 1798–1801.

- [85] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. 2015. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3460–3469.
- [86] Yan Xu, Jianwen Zhang, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. 2012. Context-constrained multiple instance learning for histopathology image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 623–630.
- [87] Yan Xu, Jun-Yan Zhu, Eric Chang, and Zhuowen Tu. 2012. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 964–971.
- [88] J Scott Yaruss, Lisa R LaSalle, and Edward G Conture. 1998. Evaluating stuttering in young children: Diagnostic data. *American Journal of Speech-Language Pathology* 7, 4 (1998), 62–76.
- [89] Hatun Zengin-Bolatkale, Edward G Conture, and Tedra A Walden. 2015. Sympathetic arousal of young children who stutter during a stressful picture naming task. *Journal of fluency disorders* 46 (2015), 24–40.
- [90] Hatun Zengin-Bolatkale, Edward G Conture, Tedra A Walden, and Robin M Jones. 2018. Sympathetic arousal as a marker of chronicity in childhood stuttering. *Developmental neuropsychology* 43, 2 (2018), 135–151.
- [91] Jing Zhai and Armando Barreto. 2006. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *2006 international conference of the IEEE engineering in medicine and biology society*. IEEE, 1355–1358.
- [92] Jing Zhai and ARMANDO Barreto. 2008. Stress detection in computer users through non-invasive monitoring of physiological signals. *Blood* 5, 0 (2008).
- [93] Han Zhang, Rui Zhang, Feiping Nie, and Xuelong Li. 2018. A generalized uncorrelated ridge regression with nonnegative labels for unsupervised feature selection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2781–2785.
- [94] Shichao Zhang, Debo Cheng, Rongyao Hu, and Zhenyun Deng. 2018. Supervised feature selection algorithm via discriminative ridge regression. *World Wide Web* 21, 6 (2018), 1545–1562.

## 8 APPENDIX

### 8.1 Presented Change-score Feature's Efficacy

This section compares our vector distance (i.e., cosine similarity and euclidean distance) based change-score feature's efficacy with conventional subtraction-based change-score features utilized in literature [17, 19, 37]. The conventional change-score for a variable ( $\Delta_f$ ) at timestamp  $t$  is calculated as the difference of the variable's value on that timestamp ( $f_t$ ) from its average at the baseline condition:  $\Delta_f = f_t - f_{baselineMean}$ . The evaluation results utilizing conventional change-score features ( $\Delta_f$ ) of HR, EDA, RSP-amp, and RSP-rate are shown in Table 7, where results are relatively poor. Compared to the conventional subtraction based change-score, cosine similarity and the euclidean distance are effective measures for vector differences (i.e., our features from each timestamp is vectors) [26, 44, 76, 80], capable of subtle difference assessment from multi-dimensional feature vectors, resulting in higher performances (Tables 1, 2, and 4).

Table 7. Evaluation of conventional  $\Delta$ -based change score features

Model	Accuracy	F1	precision	Recall	specificity
MI-MIL (Window-1 scripted)	0.50	0.61	0.50	0.79	0.33
MI-MIL (Window-2 scripted)	0.57	0.65	0.54	0.82	0.33
MI-MIL (Free speech)	0.49	0.47	0.40	0.41	0.23

### 8.2 Feature Indexes Used in Shapley Interpretability Visualization

The table 8 lists the features and their respective indices.

### 8.3 Extension of MI-MIL to Other Physiological Datasets

WESAD dataset [65] contains physiological parameter data (RSP, ECG, EDA) from 15 participants (mean age: 27.5  $\pm$  2.4 years) for baseline and stress conditions collected using a chest-worn RespiBAN device[65]. During the stress condition, the participants performed a Tier Social Stress Test (TSST) [41]. HR and Respiration features

Table 8. Feature indexes

(a) Raw physiological features				(b) Change score features			
Index	Feature name	Index	Feature name	Index	Feature name	Index	Feature name
0	Heart rate Mean	12	RSP amplitude Mean	0	Heart rate vector Cosine similarity	0	Heart rate vector Cosine similarity
1	Heart rate Min	13	RSP amplitude Min	1	Heart rate vector Euclidean distance	1	Heart rate vector Euclidean distance
2	Heart rate Max	14	RSP amplitude Max	2	EDA vector Cosine similarity	2	EDA vector Cosine similarity
3	Heart rate Median	15	RSP amplitude Median	3	EDA vector Euclidean distance	3	EDA vector Euclidean distance
4	Heart rate Var	16	RSP amplitude Var	4	RSP amplitude vector Cosine similarity	4	RSP amplitude vector Cosine similarity
5	Heart rate Std	17	RSP amplitude Std	5	RSP amplitude vector Euclidean distance	5	RSP amplitude vector Euclidean distance
6	EDA Mean	18	RSP rate Mean	6	RSP rate vector Cosine similarity	6	RSP rate vector Cosine similarity
7	EDA Min	19	RSP rate Min	7	RSP rate vector Euclidean distance	7	RSP rate vector Euclidean distance
8	EDA Max	20	RSP rate Max				
9	EDA Median	21	RSP rate Median				
10	EDA Var	22	RSP rate Var				
11	EDA Std	23	RSP rate Std				

(RSP-rate and RSP-amp) were extracted from raw ECG and Respiration effort signals. The same set of 24 statistical quantifiers (similar to the ones discussed in section 2.3.1) were extracted from 20s overlapping windows (15s overlap). The table 9 illustrates the performance of various models on the WESAD dataset. The MI-MIL approach outperformed the baseline DNN, CNN, LSTM, and Attention-based MIL models by achieving an F1 score of 0.92. Among the baseline models, the Attention-MIL achieves the second-highest F1 score of 0.91; other baseline models also achieve high accuracy (in the range of 0.85–0.90). These results also advocate the success of weakly supervised approaches as both Attention-MIL and MI-MIL perform the classification task with high accuracy.

Table 9. Evaluation baseline models and MI-MIL on WESAD dataset

Model	Accuracy %	F1 %	precision %	Recall %	ROC-AUC	specificity %
DNN	85	83	74	94	0.87	80
CNN	90	87	84	91	0.91	90
LSTM	88	84	82	87	0.88	89
LSTM (with attention)	89	86	84	88	0.89	90
Attention-MIL	91	91	91	91	0.91	91
MI-MIL	92	92	98	87	0.92	98

#### 8.4 Q4 Interpretation: MI-MIL model to Differentiate CWS vs. CWNS During Spontaneous Narration: Free-speech Dataset

This section discusses and demonstrates how MI-MIL’s inferences can be utilized and visualized to understand stuttering children’s distinctive psychological responses during spontaneous narration. The interpretations and corresponding observations at a group level are discussed in the following section.

**8.4.1 Group-wise Global Feature Importance.** The figures 5a and 5b show the global features importance through Shapley interpretation graphs of the MI-MIL models (section 4.1.4) evaluated on the free speech dataset. Each of the figures comprises two SHAP plots: features’ importance plot for true negative (i.e., detecting CWS) and true positive classification (i.e., detecting CWNS). The free speech data comprises children’s physiology during a spontaneous narration, and the evaluation in section 4.1.4 indicates that the CWS and CWNS show significant differences. This section visualizes and discusses the MI-MIL identified group-wise differences in CWS vs. CWNS through the respective model’s interpretation.

**EDA Features Importance:** As shown in figure 8a both CWS and CWNS show low EDA feature values (indicating that they were not experiencing emotional stress or arousal. Specifically, the CWS show a lower EDA mean compared to CWNS thus EDA mean acts as an important differentiating feature for the classification task. Other

important EDA features include EDA minimum and median which have high feature attribution (red blocks for stuttering class).

*HR Features Importance:* According to the figure 8a, the only HR feature that differentiated CWS and CWNS during this task was HR minimum and variance. HR variance was significantly higher in CWS than in the CWNS. Both features were important feature in differentiating CWS (red grids in X-axis index 1 and 4) vs. CWNS (blue grids in X-axis index 1 and 4). However, other HR features were not different between the groups according to the ML model and pushed the classification towards the non-stuttering class. Thus, we interpret this data to suggest that there were certain time windows where the CWS showed a higher variation in the heart rate which the model recognised as important for the classification (dark red cells index 0-5).

*RSP Features Importance:* CWNS showed slightly lower RSP-amplitude maximum and significantly higher RSP-rate maximum (feature index 20) compared to the CWS. The data for both groups are in line with speech breathing rates for preschool-age children [12]. It can be observed from the graph that the RSP-rate maximum has high contribution in classification of CWS class (red grids for feature 20). The difference in the magnitude of the RSP-rate and RSP-amp features helps that classifier in CWNS vs CWS classification (CWNS graphs have blue grids for feature 12,13,15 while CWS have red grids for feature 13,14,15,20). These results suggest that the RSP features have high contributions in the classification task for the spontaneous free speech task.

*Change-score Features Importance:* The figure 8b shows the change-score features' importance plots. According to the figures, CWS showed lower increases from baseline in almost all physiologic features, namely HR, EDA, RSP-amplitude, and RSP-rate, compared to those of CWNS. This suggests that CWNS experienced a higher arousal during the free speech task compared to CWS. The change score features were curated for the model to capture changes HR, EDA, RSP-amplitude and RSP-rate feature vectors from baseline in-terms of magnitude and direction. The euclidean distance and cosine similarity values of EDA, HR and RSP-rate feature vector were important for identifying CWS during the free speech task but these features contributed differently in different time steps as shown by the heatmap which are captured using the modality specific embedding and cross modality fusion architecture in the MI-MIL model.

According to this section's MI-MIL model interpretations for the free speech dataset, We observe that unlike the scripted dataset where CWS experienced a higher arousal during the scripted stress-provoking speech task compared to CWNS. For the spontaneous speaking task, CWNS showed higher physiologic values than CWS during this task. Lastly, both CWS and CWNS experienced higher arousal during the scripted stress-provoking speech task than during the free speech task. .

## 8.5 MIL

This section discusses the *permutation-invariance property*: An implication that can be drawn from this relationship is that the order of the bag instances does not decide the corresponding probability of the bag. This assumption further motivates MIL use since the salient CWS affective state indicative signal patterns can be independent and sparse.

*MIL Decomposition:* The MIL model tries to predict the bag label using the equation 3.

$$\theta(X) = g(\sigma_{x \in X} f(x)) \quad (3)$$

Here,  $\theta(X) \in [0, 1]$ ,  $X$  is the set of instances in a bag,  $f$  is the transformation function,  $\sigma$  is the MIL pooling function, and  $g$  is the classifier function. The choice of functions  $f$ ,  $g$ , and  $\sigma$  decide the specific approach to modeling the bag label probability. They are discussed below:

- (1) *Transformation function:*  $f$  maps the instances  $(x_j, j = 1, \dots, k)$  to low-dimensional instance-level representations  $(e_j)$  [25]. This function is applied on each individual instance belonging to the bag.

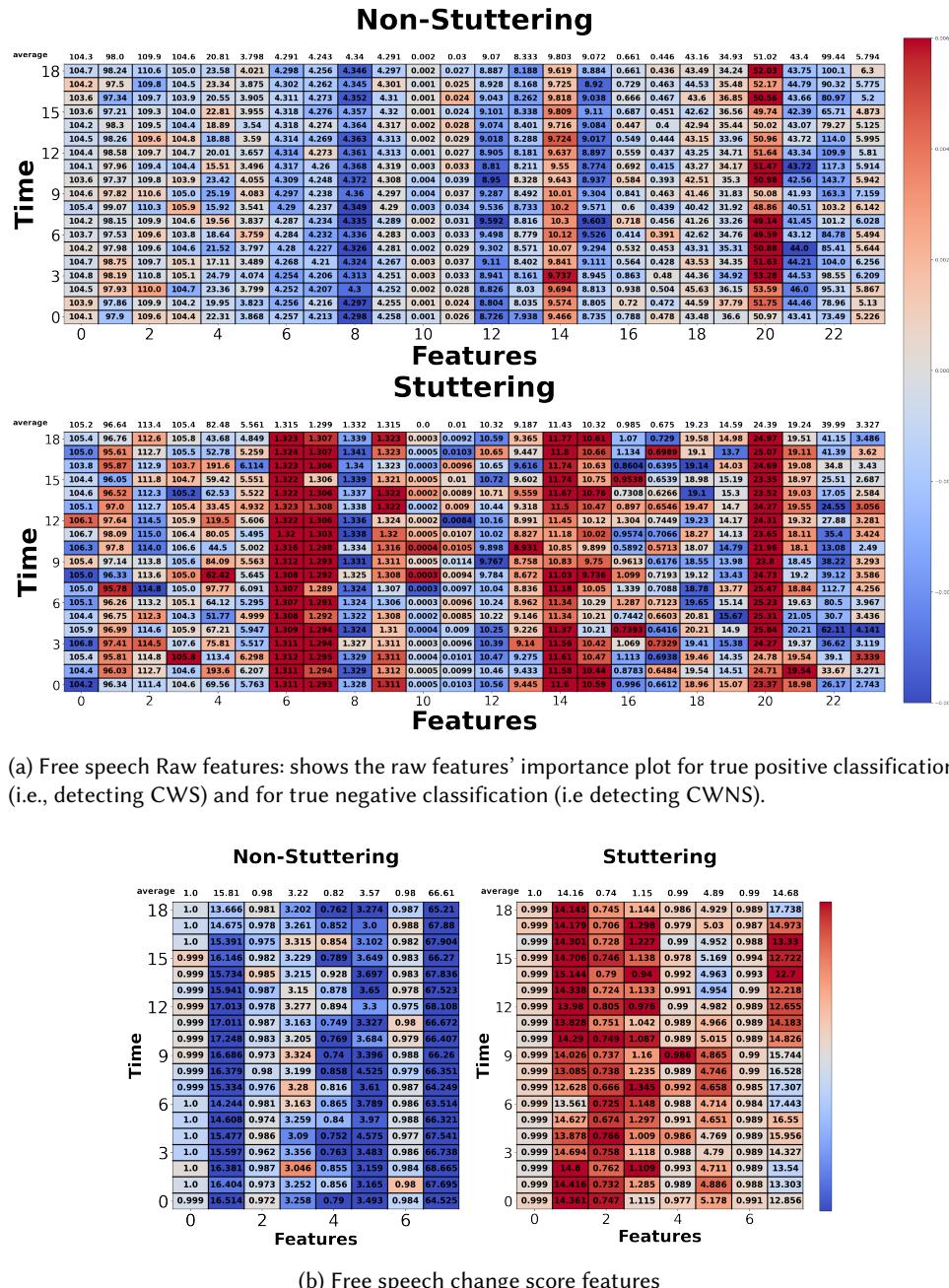


Fig. 8. Shapley plots for models evaluated on free speech data

- (2) *MIL pooling function*: Symmetric function  $\sigma$  takes the instance-level representations ( $e_j, j = 1, \dots, k$ ) as input to generate the bag-level representation. The  $\sigma$  is hence also called the aggregation function.
- (3) *Classification function*: The function  $g$  is only used in embedding-based MIL, discussed below. It takes the bag-level representation as input to generate a bag label probability score. It is used to pursue the final label of the bag.

There are two variety of MIL approaches in literature [14, 35]:

*Instance-Based MIL Approach*: It is the common form of MIL evaluated in the literature. A *transformation function*  $f$  takes each instance as input  $x_j$  and returns the one-dimensional instance-level scores (i.e., instance labels  $e_j = y_j$ ). Then individual instance-level scores are aggregated (through pooling) to obtain the bag label  $Y$ . Max or mean operators are generally used as the pooling functions. These two functions are symmetric and do not violate the permutation invariant assumption of the MIL approach. Studies have used other functions like convex maximum [38, 60] as the pooling function  $\sigma$ .

*Embedding-Based MIL Approach*: In this approach, a *transformation function*  $f$  maps the instances ( $x_j, j = 1, \dots, k$ ) to a lower  $m$ -dimensional embeddings ( $e_j$ ).  $m$  is a hyper-parameter. A MIL pooling  $\sigma$  takes the embeddings and generates a bag representation  $z$  that is independent of the number of instances in the bag. A classifier function  $g$  further process the bag representation to infer the bag label  $Y$ . This approach is arbitrarily flexible and can be trained by backpropagation. The only constraint is that the function  $\sigma$  must be differentiable [35, 81].

Previous studies have shown that the embedding-based MIL approaches achieve better classification performance [81]. Since the individual instance labels are unknown, the transformation functions  $f$  may be trained insufficiently and introduce error to the bag-level class prediction. The embedding-based approaches generate a joint representation of a bag from the instances; hence they do not introduce any additional bias to the bag-level classification [35, 81]

### 8.5.1 MI-MIL Implementation.

As discussed in section 2.3.1,  $19 \times 24$  raw physiological features are extracted from the 20s signal. There are 19 instances, and each instance (i.e., 2s signal) is represented by a  $1 \times 24$ -dimensional feature set.

The MI-MIL architecture for the physiological features evaluated on the scripted dataset and free speech dataset is as follow: The MI-MIL implementation consisted of 4 separate modality specific embedding blocks. The embedding block for each modality consisted of a linear layer [128] with a ReLU activation function. The output from the linear layer is flattened and passed through a 10 % dropout layer which is followed by linear layer [256]. After generating a modality specific embedding we pass the embedding to the self attention block. The modality specific attention block consists of two linear layer[256, 1] which are separated by a *tanh* function. The modality fusion block consists of three Conv1d layers with 4 input channels and 2 output channels and a (kernel size=1,stride=1,padding=0). Since we concatenate embeddings from four modalities we get a combined bag representation of the dimensions <4,256> which is passed to the modality. The output from the modality fusion block goes through classification block which consists of three linear layers [256, 64, 1] with ReLU activation between the first two layers and sigmoid activation function for the last layer of the classifier. The figure 9 illustrates the layer wise structure for different blocks.

We extract  $19 \times 8$  change-score physiological features from each 20s signal. Thus, there are 19 instances, and each instance (i.e., 2s signal) is represented by a  $1 \times 8$ -dimensional change-score vector. The Architecture of the change score MI-MIL model is same as the scripted dataset apart from the dropout layer (20% dropout is used) in the HR and RSP rate modality embedding block.

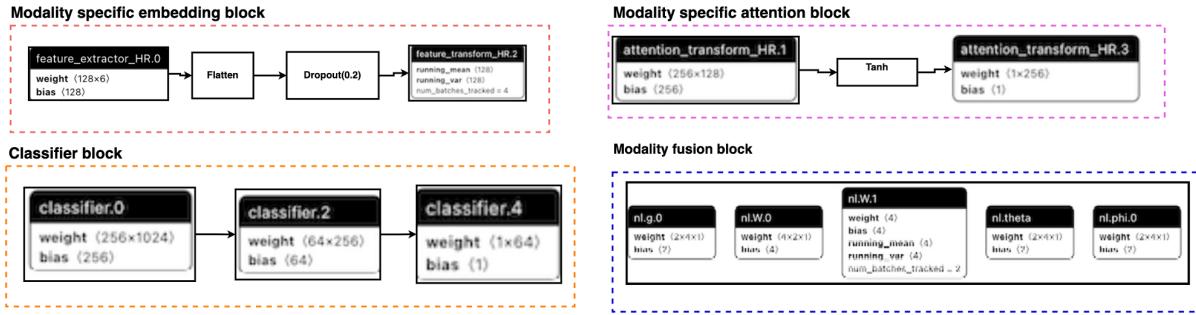


Fig. 9. Layer-wise architectural information for MI-MIL

**8.5.2 Attention-MIL Implementation.** The attention-MIL architecture for the raw physiological features evaluated on the scripted dataset is as follow: The embedding block consists of a linear layer [200], a 1-D convolution layer with 20 kernel and 2X1 kernel size, and another linear layer [64] with ReLU activation function in each layer. A flatten layer is used to make the output two-dimensional, followed by a batch normalization layer and a linear layer [128] with ReLU activation. The output of the embedding block is fed into the attention block consisting of two linear layer[64, 1] which are separated by a *tanh* function (figure ??). The classification block consists of two linear layers [64, 1] with ReLU and sigmoid activation functions.

The attention-MIL architecture for the raw physiological features evaluated on the free-speech dataset is as follow: Since detecting affective states difference in CWS vs. CWNS from spontaneous narration is a harder task, and the free-speech data sample size is larger, the transformation/embedding block is more complex in this network. The transformation function block consists of four 1D-convolution layers [256, 512, 128, 32] and a kernel size of 2X1 with ReLU activation. After flattening the output from the convolution layers, 1D batch normalization was used, followed by two linear layers [512, 128] with ReLU activations. The attention and classifier blocks were the same as the model evaluated on the scripted dataset discussed above.

The attention-MIL architecture for the change-score features evaluated on the scripted dataset is as follow: The transformation function/embedding block has two 1-D convolution layers [128, 64] followed by two linear layers [256, 256] and an intermediate 10% dropout layer between the linear layers. Each of these layers uses a ReLU activation. Akin to the raw features model, the change score model flattens the output and feeds it into a 1D batch normalization, followed by a 10% dropout layer and two linear layers [512, 256] with ReLU activation present after each of the layers. The attention block consisted of two linear layers [64, 1] with an intermediate *tanh* function. The structure of the classification block consisted of four linear layers [200, 150, 64, 1]. All linear layers had ReLU activation apart from the final layer, which had a sigmoid activation function.

The attention-MIL architecture for the change-score features evaluated on the free-speech dataset is as follow: The transformation function or embedding block has two convolution layers [64, 32] with a kernel size of 2X1 trailed by two linear layers [128, 16]. Each of the convolution and linear layers have a ReLU activation function. The output is flattened, fed into a 1D batch normalization layer, and fed into two linear layers [1024, 512]. All the linear and convolution layers have ReLU activation functions. The attention block contains two linear layers [64, 1] separated by a *tanh*. The structure of the classification block remained the same as the raw features model evaluated on the scripted dataset discussed above.

**8.5.3 Baseline Model Implementation.** In each of the evaluations, we considered four supervised learning classifiers: long short-term memory (LSTM), convolutional network (CNN), deep neural network (DNN), and LSTM

with self-attention network architectures as baseline models. The baseline model implementations are discussed below:

*LSTM model* was evaluated to explore the sequential dependence among the physiological data. The implementation consists of a single-layered LSTM and a linear output layer with Sigmoid activation.

The *CNN model* can extract the complex sequential and global information of the data. It comprises 1D-convolution layers with kernel size 2X1 and padding size of 1. A 20% dropout is used between the convolution layers, and ReLU is used as the activation function. The output is then flattened and passed through some linear layers followed by 1D batch normalization layers. Each intermediate layer has ReLU as the activation function with Sigmoid activation in the output layer.

The *DNN model* architecture consists of some Linear layers, which are flattened and passed through the 1D batch normalization layer. After batch normalization, the outputs are passed through a dense layer with a sigmoid activation function.

In order to see the performance of supervised learning approaches using the attention framework. We evaluated the *LSTM model (with self attention)*. The model architecture consists of a four-layer LSTM whose output is fed to a self-attention block (similar to the one discussed in 3.2.2) which consists of two linear layers separated by a *tanh* activation. The attention block generates the attention weight for each instance after receiving the embedding from the LSTM layer. Then, the attention-weights are multiplied with the embedding generated from the LSTM model (self-attention), which is passed through a linear layer [128] and an output layer.

## 8.6 Overview of the existing deep learning literature on stress detection

The table 10 lists the existing work in deep learning domain which are focused on stress detection.

Table 10. Overview of the existing literature on Stress detection (Abbrev.- BVP:Blood Volume Pulse, BR: Breathing rate, EDA: Electrodermal activity,FDA:Fisher Discriminant Analysis, GSR: Galvanic skin response,HR: Heart rate,PD: Pupil Diameter,RESP: Respiration activity, ST: Skin Temperature, SVM: Support Vector Machine, KNN: k- nearest neighbours, CNN: Convolutional Neural Network)

Stress Type	Sensor data	Literature	Age-group	Method
Paced Stroop Test	GSR, BVP, ST, PD	[91]	21-42	SVM
Spielberger stress measurement questionnaire Memory test	EDA, HR	[27]	59.8+-5.8	Pvalue and Correlation analysis
Talk Preparation, Hyperventilation	HR, EDA	[20]	19-32	Fuzzy Logic
Mental Arithmetic Task, , logical puzzle task	ECG, EDA,RESP EMG	[84]	19-53	Bayes Normal , Quadratic Bayes Normal, K-NN , Fisher's Least Square Linear Classifier
Controlled Trier social stress test (TSST)	EDA, PPG, HR	[57]	18-39	SVM,AdaBoost, KNN
Paced Stroop Test	EDA, BVP, ST, PD	[92]	21-42	SVM Naives Bayes, Decision Tree
Stroop Test	PPG, ECG, PD	[56]	22-28	SVM Genetic Algorithm, Fuzzy SVM
Deception	GSR, BVP, ST, BR,	[3]	20-35	Decision Tree Classifier
Talk Preparation, Hyperventilation	HR, GSR	[21]	19-32	FDA, KNN Classifier
Applicants played custom version of game: tetris	BVP, HR, EDA	[53]	18-32	CNN
TSST	RESP, ECG, EDA, EMG, ST	[34]	25-30	CNN
Scary Video, Puzzle Task, Pain, Physical Activity	EDA, BVP	[4]	25-30	Decision Tree, RNN, RCNN