Non Contact Stress Assessment Based on Deep Tabular Method

Urmila[a] and Avantika Singh[b]
Dr. Shyama Prasad Mukherjee International Institute of Technology, Naya Raipur, Raipur, Chhattisgarh, India
{*urmila23300, avantika*}*@iiitnr.edu.in*

Keywords: Stress Detection, Remote Photoplethysmography (rPPG), Physiological Features.

Abstract: In today's competitive world, stress is the major factor that influences human health negatively. In the long term, stress can lead to serious health problems such as diabetes, depression, anxiety, and various heart diseases. Thus, timely stress recognition is important for efficient stress management. Currently, for stress assessment various wearable devices are used to capture physiological signals. However, these devices although accurate are cost-sensitive and requires direct physical contact which may lead to discomfort in long run. In this work we have introduced a tabular based deep learning architecture for detecting stress by analyzing physiological features. The architecture extracts physiological features from remote photoplethysmography (rPPG) signals computed from facial videos. The proposed architecture is validated on publicly available UBFC-Phys dataset for two sets of experiments (i) Stress task classification and (ii) Multi-level stress classification. For both set of experiments the proposed methodology outperforms the current state-of-art method. The code is available at `https://github.com/Heeya2205/Deep_tabular_methods`.

1 INTRODUCTION

Stress is a physical, emotional, or mental response to external pressures or demands, which can arise from various factors. Life situations like work challenges, family issues, and environmental changes often arouse stress conditions in an individual. It is a natural response of the body to perceived threats and can manifest as tension, anxiety or other psychological or physical reactions (Giannakakis et al., 2019). Stress causes the body to release hormones like cortisol and adrenaline, which raise your heart rate, blood pressure, and blood sugar to help you deal with challenges. However, if stress lasts a long time, it can lead to serious health problems like heart disease, diabetes, anxiety, and depression (Fink, 2010).

In human body autonomic nervous system (ANS) is responsible for controlling the number of heartbeats per minute (Barazi et al., 2021). ANS can be further divided into two sub-parts: (a) Sympathetic nervous system, (b) Parasympathetic nervous system. Sympathetic nervous system is active when an individual encounters situation like stress, anxiety, fear or undergoes laborious exercises, as a result it increases the heart rate. On the other hand parasympathetic nervous system is active when an individual is in calm state of mind particularly when an individual feels compassion or love and thus, it decreases the heart-rate. As mentioned above heart rate is directly influenced by ANS which further depends upon our state-of-mind. Thus in this work we aim to estimate stress based on heart-rate (HR) and its related factors like Heart rate variability (HRV), Peak detection etc.

Traditionally, heart rate can be measured by various methods like electrocardiography (ECG), electromyography (EMG), and photoplethysmography (PPG). However, these methods although accurate are cost-sensitive and requires direct physical contact. Henceforth, in this work we explore, non-invasive method of heart-rate estimation by utilizing remote photoplethysmography (rPPG) signals.

Here, in this work for stress assessment, firstly rPPG signals are extracted from recorded video frames. Later, physiological features like HR, HRV etc are calculated from observed rPPG signals. Further, the extracted physiological features are encoded for learning discriminative stress levels by employing deep tabular data learning architecture (Arik and Pfister, 2021) based on attention mechanism. Main contributions of this work are as follows:

- **Efficient Tabular Learning with TabNet:** The proposed method harnesses TabNet network (Arik and Pfister, 2021) as the foundational backbone to directly process physiological features ex-

[a] https://orcid.org/0009-0003-8642-8525
[b] https://orcid.org/0000-0002-2606-5959

Urmila, and Singh, A.
Non Contact Stress Assessment Based on Deep Tabular Method.
DOI: 10.5220/0013150300003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 597-604
ISBN: 978-989-758-730-6; ISSN: 2184-4313

tracted from rPPG signals, incorporating built-in feature prioritization and sparsity mechanisms. This eliminates the need for extensive preprocessing while enhancing model interpretability and performance.

- **Improved Generalization Across Classification Tasks:** The method demonstrates lower variance in accuracy across binary and multi-class stress classification tasks, highlighting superior generalization compared to existing approaches, which often exhibit significant performance variability.
- **Robust Feature Extraction:** To effectively manage the dynamic motion of faces across video frames and improve the precision of facial region extraction, the proposed method integrates the Haar Cascade (Choi et al.,), and Mediapipe library (Lugaresi et al., 2019).This combination facilitates robust face detection and accurate localization of facial landmarks.

2 RELATED WORK

Several works have been reported in the literature for rPPG signal analysis (Das et al., 2023) (Speth et al., 2024) but the work done in the field of stress analysis using rPPG signal is still limited. In one of the notable work (Ziaratnia et al., 2024) author's proposed deep learning-based method that utilize Compact Convolutional Transformers (CCT) for feature extraction and Long Short-Term Memory (LSTM) for temporal pattern recognition. This study presents a non-contact approach for recognizing stress by utilizing rPPG signals derived from RGB facial videos. In another work (Xu et al., 2024) a multi-task attentional convolutional neural network (MTASR) is deployed, which integrates peak detection and heart rate estimation to enhance stress recognition. By employing peak detection as a physiological parameter, the researchers trained their network to identify more reliable physiological indicators. In another work authors (Casado et al., 2023) recognized depression based on a pipeline that extract rPPG signals in a full unsupervised manner, and calculate 60 statistical, geometrical, and physiological features. These extracted features are further used to train several machine learning regressors to recognize different level of depressions. In (Ntalampiras, 2023) authors proposed a non-intrusive, low-cost, and automatic stress monitoring framework. This framework extracts multi-domain speech features to reveal complementary stress-related characteristics. In (Pan et al., 2024) a deep network for stress assessment is proposed that focuses on facial features such as expressions, movements, and specific areas like the eyes, nasolabial folds, and jaw, which are related to depression. The proposed model emphasizes dynamic facial features through its attention mechanism.

3 METHODOLOGY

This section conceptualizes our proposed approach. Figure 1 gives a generic overview of the proposed system which consists of mainly five parts: (i) ROI selection and facial landmark detection (ii) Color space extraction from cropped facial regions (iii) rPPG signal extraction (iv) Physiological features extraction from rPPG signals (v) Tabular learning on physiological features.

3.1 ROI Selection and Facial Landmark Detection

In the pre-processing phase, video frames are extracted from facial video sequences. Specifically, we utilize a frame rate of 30 fps for temporal segmentation of the video. Each extracted frame is maintained at a resolution of 513×513 pixels. Following the frame extraction, the next step involves generating a bounding box around the facial region of interest. To achieve this, we implement the Haar Cascade algorithm (Choi et al.,), which operates by detecting facial features through the application of rectangular filters that compute the intensity differences between adjacent regions in the image. Once the bounding box is delineated, the subsequent step is the detection of facial landmarks. For this, we employ the Mediapipe Face Mesh (Lugaresi et al., 2019) framework, which facilitates the identification of 478 facial landmarks (Ziaratnia et al., 2024).

By integrating both techniques in our proposed method, we address the challenges inherent in extracting the facial region of interest. Figures 2a and 2b demonstrate these challenges and highlight the improvements achieved through our approach.

Furthermore, convex hull based masking technique has been employed to isolate and emphasize the facial region. Mathematically, the convex hull is the smallest convex shape that can enclose all the points representing the facial landmarks. For this we have first created a binary mask with the same resolution as our extracted video frame resolution (513×513). Formally, binary mask is described as:

$$\text{mask}(x,y) = \begin{cases} 255 & \text{if } (x,y) \in \text{convex hull} \\ 0 & \text{otherwise} \end{cases}$$

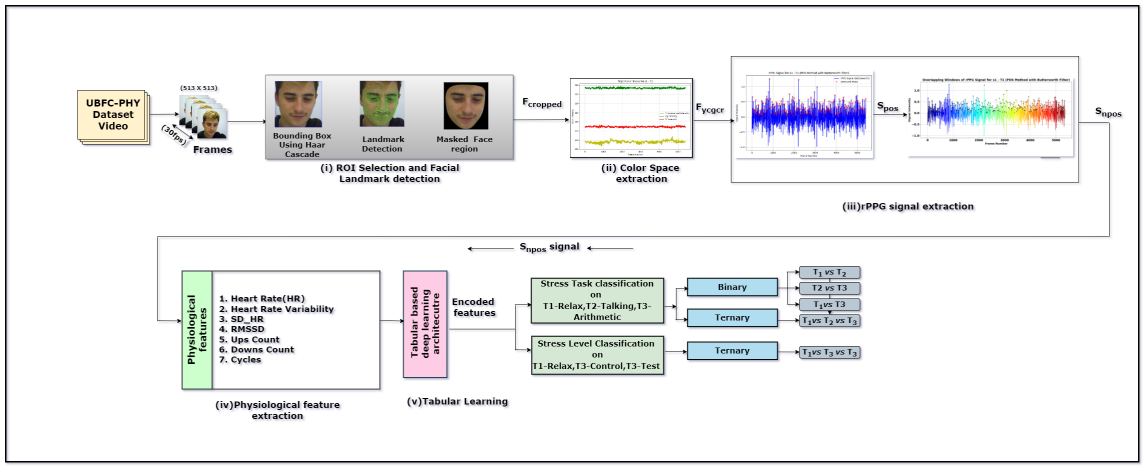


Figure 1: Proposed architecture comprising of five main parts: (i) ROI selection and facial landmark detection (ii) Color space extraction (iii) rPPG signal extraction (iv) Physiological features extraction (v) Tabular learning on extracted physiological features.

where (x, y) denotes the pixel coordinates. The last step is to generate cropped facial region, which is generated as:

$$F_{Cropped} = F_{Mediapipe} \wedge \text{mask}$$

where, $F_{Mediapipe}$ is the output representing 478 facial landmarks. Here, the bitwise AND operation is used to retain only the pixel values of the facial region that fall within the convex hull, effectively setting all non-facial pixels to black. This selective masking technique highlights the facial features by suppressing irrelevant background elements.

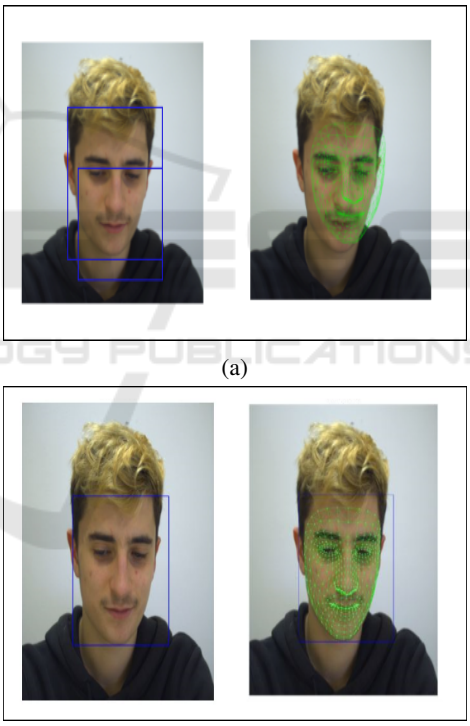
3.2 Color Space Extraction from Cropped Facial Regions

Inspired by the prior study (Kim et al., 2021) we apply YCgCr color space (F_{YCgCr}) from $F_{Cropped}$ (cropped facial region). By successfully decoupling the luminance component (Y) from the chrominance components (Cg) and (Cr), this color space facilitates the identification of subtle color changes that signify physiological conditions. By concentrating on the luminance channel and working with various skin tones, this color space lessens lighting variations (Panigrahi and Sharma, 2022).

3.3 rPPG Signal Extraction

For extracting rPPG signals from F_{YCgCr} we have utilized state-of-the-art POS (Plane-Orthogonal-to-Skin) (Wang et al., 2016) algorithm. Algorithm 1 illustrates POS detailed steps.

Inspired by the prior study (Xu et al., 2024) to normalize extracted rPPG signals (S_{pos}) we have filtered it with Butterworth filter (Selesnick and Burrus, 1998). While filtering low cut-off frequency was set as 0.7 Hz and high cut-off frequency was set as 2.5 Hz. Once the signals are normalized, we divide them into 2-second windows (30 frames per second) with a 10% overlap between segments to account for temporal differences across frames.



(a)

(b)

Figure 2: Preprocessing steps. (a) Without proper preprocessing . (b) After preprocessing. The image is taken from UBFC-Phys dataset.

Algorithm 1: rPPG Signal Extraction.

Input: F_{YCgCr}
Output: S_{pos} (extracted rPPG Signal)
Step 1: Compute Color Signals
$C_1 \leftarrow Y - Cg$ ▷ Color Signal 1
$C_2 \leftarrow Cr - \frac{Y+Cg}{2}$ ▷ Color Signal 2
Step 2: Compute Means of Color Signals
$\mu_1 \leftarrow \frac{1}{n}\sum_{i=1}^{n} C_{1i}$ ▷ Mean of Color Signal 1
$\mu_2 \leftarrow \frac{1}{n}\sum_{i=1}^{n} C_{2i}$ ▷ Mean of Color Signal 2
Step 3: Adjust Color Signals by Subtracting the Mean
$S_1 \leftarrow C_1 - \mu_1$ ▷ Adjusted Signal 1
$S_2 \leftarrow C_2 - \mu_2$ ▷ Adjusted Signal 2
Step 4: Compute Final rPPG Signal
$S_{\text{pos}} \leftarrow S_1 + S_2$ ▷ rPPG Signal
Return S_{pos}

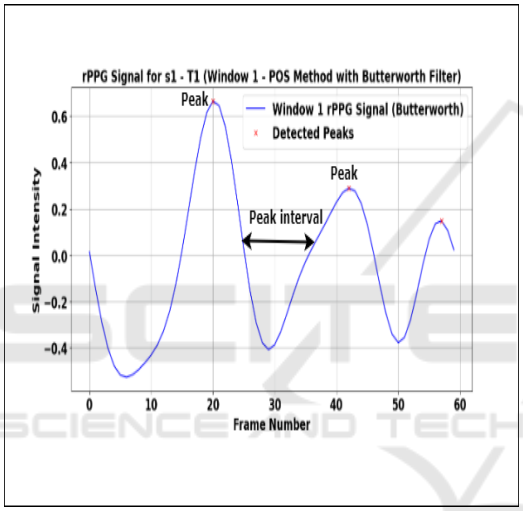


Figure 3: Single window depicting signal peak and interval detection.

3.4 Physiological Features Extraction

From the normalized rPPG signals (S_{npos}) we have extracted signal peaks as shown in Figure 3. From the extracted peaks, signal peak intervals are computed as depicted in Figure 3. Later on the basis of aforementioned signal intervals 7 physiological features are computed as follows:

(i) Heart Rate (HR):

$$HR = \frac{60}{\text{mean interval between peaks (seconds)}}$$

where intervals are the time differences between consecutive S_{npos} peaks.

(ii) Heart Rate Variability (HRV):

$HRV =$ Standard deviation of intervals (seconds)

which represents the variation in time between heartbeats.

(iii) Standard Deviation of Heart Rate (SD_HR):

$SD\_HR =$ Standard deviation of computed heart rates (bpm) calculated for each interval.

(iv) Root Mean Square of Successive Differences (RMSSD):

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(\text{interval}_{i+1} - \text{interval}_i)^2}$$

where N is the number of intervals, indicating short-term HRV.

(v) Ups Count: In a sympathetic activity in Autonomic nervous system(ANS) this ups count refer to the number of times S_{npos} signal transit from a lower value to higher value. It indicates the onset of a heartbeat.

(vi) Downs Count: It is the moment when the heart finishes pumping blood and start to relax, which is captured as a downward transition in the S_{npos} signal.

(vii) Cycles: Mathematically, it is defined as the minimum of Ups Count and Downs Count.

3.5 Tabular Learning on Physiological Features

The extracted physiological features are numeric values; thus, we store them in a tabular structure. To extract meaningful information from it, we have incorporated the state-of-the-art tabular learning architecture, TabNet (Arik and Pfister, 2021). In our proposed methodology, we have extracted discriminative information from physiological features by utilizing the TabNet encoder architecture, as depicted in Figure 4. This tabular-based deep learning architecture operates through a series of sequential decision steps, where each step refines its focus based on the information processed from the previous step. Each decision step employs non-linear transformations to refine the feature representations. Furthermore, it includes a sparsity regularization mechanism to control the number of features selected at each step.

As depicted in Figure 4, this architecture mainly comprises three modules: (i) Feature Transformer, (ii) Attentive Transformer, and (iii) Feature Selection Mask. All these are described below:

Feature Transformer: This module is responsible for transforming the input physiological features into meaningful representations. These representations are later divided into two parts using a split block. The first part, denoted as $d[i]$ (as depicted in Figure 4), contributes to the current decision output,

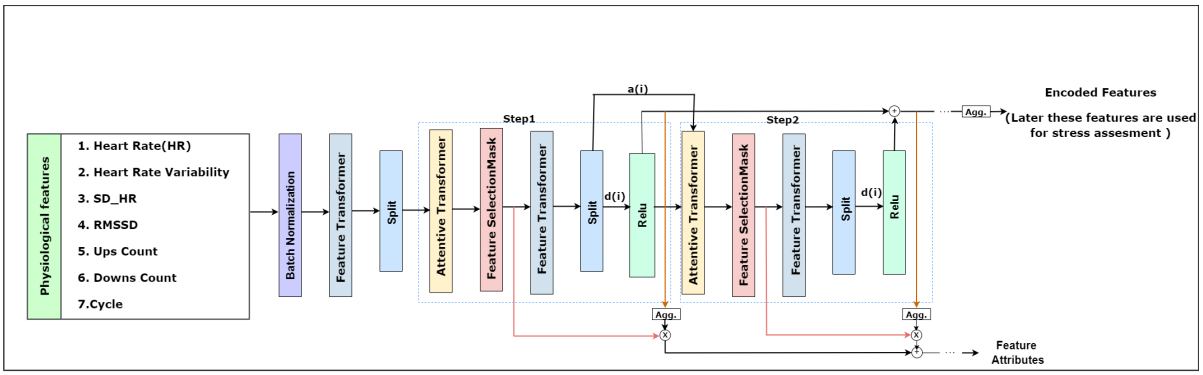


Figure 4: Encoder architecture for extracting discriminative information from physiological features. This architecture operates through a series of sequential decision steps mainly comprising of three parts (i) Feature Transformer (ii) Attentive Transformer (iii) Feature Selection Mask. Here, $a[i]$ serves as the input for the attention transformer in the next decision step while $d[i]$, denotes the current decision output.

while the other part, denoted as $a[i]$ (as depicted in Figure 4), serves as the input for the attentive transformer in the next decision step.

Attentive Transformer: This module generates a trainable mask that identifies the most important features for each decision step using a sparsemax function, which creates a sparse, interpretable feature selection mechanism. A prior scale term is also employed to regulate how often each feature is selected across decision steps.

Feature Selection Mask: This module is used for selecting globally important feature attributions.

To develop a deep tabular learning model capable of robustly identifying and prioritizing specific features derived from rPPG signals that play a critical role in our analysis, we leverage the TabNet architecture. This approach eliminates the need for preprocessing techniques by allowing the model to adaptively focus on essential features. If a feature is deemed less supportive at any stage, the model advances to another level of evaluation to reassess its importance.

4 EXPERIMENT AND RESULTS

4.1 Dataset Description

To validate the performance of proposed approach in this work we have worked on UBFC-Phy dataset (Sabour et al., 2021). This dataset is publicly available and consists of 56 subjects in total. Out of these 56 subjects 47 subjects are female and rest are male. This dataset contains facial videos of enrolled subjects. Each subject underwent three different tasks: resting task (T1), speech task (T2) and arithmetic task (T3) resulting in generation of 168 videos in total. Furthermore, this dataset is divided into two groups based on the challenging nature of the task namely: control group and test group. In control group subjects faced less challenging task as compared to subjects in test group.

4.2 Model Validation and Experimental Setup

For validating our proposed approach we have conducted two set of experimentation as conducted in previous state of the art work (Ziaratnia et al., 2024) namely: (i) Stress task classification, (ii) Multi-level stress classification.

In stress task classification subjects are classified on the basis of tasks performed. Under this we have performed both binary ($T1$ vs. $T2$, $T2$ vs. $T3$, and $T1$ vs. $T3$) as well as ternary classification ($T1$ vs. $T2$ vs. $T3$). While in case of multi-level stress classification subjects are classified on the basis of the challenging nature of the task. Out of the three tasks $T3$ (arithmetic task) is considered as the most demanding task (Ziaratnia et al., 2024). Thus, this set of experimentation is focused towards classifying stress into three levels: (i) no stress ($T1$ task for both control and test group), (ii) low stress ($T3$ control group), and (iii) high stress ($T3$ test group). For stress task classification seven-fold cross-validation technique and for multi-level stress classification stratified five-fold cross-validation technique has been adopted as in (Ziaratnia et al., 2024).

All the experiments have been conducted on a system with specifications as: Intel(R) Xeon(R) Silver 4114 CPU, 64 GB of RAM, 4 TB SSD, NVIDIA GeForce GTX 1080 Ti GPU. For the software environment, we have utilized libraries such as OpenCV (Mordvintsev and Abid, 2017), PyTorch (Imambi

et al., 2021), and Mediapipie (Lugaresi et al., 2019).

Table 1: Stress task 7-fold cross-validation classification results. Binary classification **T1 (resting task) vs. T2 (speech task).**

K-folds	Accuracy	Precision	Recall	F1 score
Fold-1	0.750	0.833	0.625	0.714
Fold-2	0.681	0.615	1.000	0.761
Fold-3	0.812	0.727	1.000	0.842
Fold-4	1.000	1.000	1.000	1.000
Fold-5	0.937	0.888	1.000	0.941
Fold-6	0.937	0.888	1.000	0.941
Fold-7	0.875	1.000	0.750	0.857
7-Fold Mean	**0.857**	**0.850**	**0.9107**	**0.865**

Table 2: Stress task 7-fold cross-validation classification results. Binary classification **T1 (resting task) vs. T3 (arithmetic task).**

K-folds	Accuracy	Precision	Recall	F1 score
Fold-1	0.812	0.857	0.750	0.800
Fold-2	1.000	1.000	1.000	1.000
Fold-3	0.937	1.000	0.875	0.933
Fold-4	0.930	1.000	0.875	0.933
Fold-5	1.000	1.000	1.000	1.000
Fold-6	1.000	1.000	1.000	1.000
Fold-7	0.937	0.888	1.000	0.941
7-Fold Mean	**0.946**	**0.963**	**0.928**	**0.944**

Table 3: Stress task 7-fold cross-validation classification results. Binary classification **T2 (speech task) vs. T3 (arithmetic task).**

K-folds	Accuracy	Precision	Recall	F1 score
Fold-1	0.750	0.833	0.625	0.714
Fold-2	0.875	0.875	0.875	0.875
Fold-3	0.812	0.857	0.750	0.800
Fold-4	0.875	1.000	0.750	0.857
Fold-5	0.812	0.857	0.750	0.800
Fold-6	0.875	1.000	0.750	0.857
Fold-7	0.937	0.888	1.000	0.941
7-Fold Mean	**0.848**	**0.901**	**0.785**	**0.835**

Table 4: Stress task 7-fold cross-validation classification results. Ternary classification **T1 (resting task) vs. T2 (speech task) vs. T3 (arithmetic task).**

K-folds	Accuracy	Precision	Recall	F1 score
Fold-1	0.750	0.757	0.750	0.752
Fold-2	0.750	0.795	0.750	0.752
Fold-3	0.875	0.891	0.875	0.873
Fold-4	0.916	0.933	0.916	0.918
Fold-5	0.833	0.838	0.833	0.829
Fold-6	0.916	0.933	0.916	0.918
Fold-7	0.916	0.933	0.916	0.918
7-Fold Mean	**0.851**	**0.869**	**0.851**	**0.851**

4.3 Performance Metrics

The proposed framework's performance is evaluated using common classification evaluation criteria, including accuracy, precision, recall, and F1 score. Accuracy is calculated as the ratio of correctly predicted instances to total instances, while precision is the ratio of accurately predicted positive cases to all instances predicted as positive. Recall is the proportion of accurately predicted positive instances relative to all actual positive instances, including those incorrectly classified as negative. F1 Score is calculated as the harmonic mean of precision and recall.

Table 5: Multi-Level stress 5-fold cross-validation classification results. Ternary classification: **no stress (T1 task for both control and test group), low stress (T3 control group) and high stress (T3 test group).**

K-folds	Accuracy	Precision	Recall	F1 score
Fold-1	0.826	0.896	0.826	0.816
Fold-2	0.869	0.864	0.869	0.864
Fold-3	0.863	0.877	0.863	0.857
Fold-4	0.954	0.961	0.954	0.953
Fold-5	0.863	0.865	0.863	0.861
5-Fold Mean	**0.875**	**0.882**	**0.875**	**0.870**

Table 6: Multi-Level stress 5-fold cross-validation classification results. Mean value (± standard deviations) for three class classification individually: **no stress (T1 task for both control and test group), low stress (T3 control group) and high stress (T3 test group) .**

Class	Accuracy	Precision	Recall	F1 score
No Stress	0.981(±0.040)	0.910(±0.092)	0.981(±0.040)	0.943(±0.061)
Low Stress	0.866(±0.139)	0.812(±0.059)	0.866(±0.139)	0.835(±0.088)
High Stress	0.653(±0.086)	0.910(±0.124)	0.653(±0.086)	0.756(±0.081)

4.4 Stress Task Classification Performance

As mentioned in section 4.2 for stress task classification experimentation we have computed results for both binary as well as ternary classification. Table 1, Table 2, and Table 3 illustrates binary stress classification results for T1 (resting task) vs. T2 (speech task), T1 (resting task) vs. T3 (arithmetic task), and T2 (speech task) vs. T3 (arithmetic task) tasks respectively. Table 4 illustrates ternary stress task classification results for T1 (resting task) vs. T2 (speech task) vs. T3 (arithmetic task). Major observations from Table 1, Table 2, Table 3, and Table 4 are as follows:

- Highest accuracy is observed while differentiating subjects undergoing resting task versus arithmetic task as depicted in Table 2. As stated in (Ziaratnia et al., 2024) arithmetic task is the most challenging task and our proposed approach effectively discriminates it which depicts our model superiority.
- Lowest recall is observed while differentiating subjects undergoing speech task versus arithmetic task as depicted in Table 3. This phenomenon is quite obvious because speech task that involves expressing views in front of others and solving arithmetic problems both involved inducing stress in individuals.

Table 7: Experimental results comparing our approach to other cutting-edge techniques for stress task categorisation on the UBFC-Phys dataset. The best findings are bolded in the table, which displays the mean values (± standard deviations) of the 7 cross-validation. This table's values are derived from (Ziaratnia et al., 2024).

Methods	T1 vs. T2		T1 vs. T3		T1 vs. T2 vs. T3	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
MLP (Dolmans et al., 2021)	0.709(±0.061)	0.706(±0.064)	0.599(±0.040)	0.587(±0.037)	0.440(±0.028)	0.434(±0.031)
LIT (Dolmans et al., 2021)	0.701(±0.063)	0.699(±0.066)	0.625(±0.042)	0.622(±0.052)	0.447(±0.027)	0.443(±0.031)
DFAF (Gao et al., 2019)	0.758(±0.035)	0.756(±0.033)	0.689(±0.040)	0.686(±0.037)	0.478(±0.034)	0.477(±0.036)
CAM (Praveen et al., 2021)	0.726(±0.060)	0.722(±0.063)	0.650(±0.052)	0.645(±0.054)	0.494(±0.028)	0.487(±0.033)
MFN (Yu et al., 2021)	0.769(±0.035)	0.763(±0.043)	0.684(±0.050)	0.670(±0.049)	0.501(±0.038)	0.500(±0.036)
BCSA (Zhang et al., 2023)	0.818(±0.063)	0.817(±0.063)	0.723(±0.039)	0.722(±0.039)	0.558(±0.052)	0.552(±0.048)
Multimodal CCT-LSTM (Ziaratnia et al., 2024)	**0.981(±0.016)**	**0.981(±0.016)**	0.924(±0.037)	0.924(±0.037)	0.832(±0.058)	0.834(±0.056)
Proposed Approach	0.857(±0.104)	0.865(±0.095)	**0.946(±0.062)**	**0.943(±0.065)**	**0.851(±0.069)**	**0.851(±0.070)**

- In case of ternary stress task classification task as depicted in Table 4 the mean accuracy, precision and recall across 7-folds are almost same. This phenomenon indicates fairness of our model which focuses not only on minimizing false negatives but also focuses on minimizing false positives.

4.5 Multi-Level Stress Classification Performance

Table 5 and Table 6 depicts results for multi-level stress classification. According to state-of-the-art research, we employed a 5-fold stratified cross validation strategy in this experiment because the number of subjects in the control and test groups was unbalanced (Ziaratnia et al., 2024). As depicted in Table 5 the highest accuracy reported across all folds was in fold-4 while the mean accuracy achieved was 87.5%. Furthermore, as depicted in Table 6 the highest and lowest multi-level stress accuracy was achieved in No-Stress and High-Stress class respectively.

4.6 Comparative Analysis

To validate the efficacy of proposed approach we have compared our computed results with other state-of-the-art methods working on same dataset (UBFC-Phy) as ours. It is evident from Table 7 that, with one exception (T1 vs. T2), our suggested methodology performs better than any previous state-of-the-art effort. When compared to the most recent state-of-the-art work (Ziaratnia et al., 2024), the findings for **(i) T1 vs. T3** and **(ii) T1 vs. T2 vs. T3** demonstrate an accuracy improvement of 2.2% and 1.9%, respectively. For our proposed approach the results obtained in case of T1 vs. T2 case is less as compared to the available state-of-the-art approach. It should be noted that in case of current available state-of-the-art work (Ziaratnia et al., 2024) there is variation of about 15% in accuracy (as reported in Table 7) while considering various stress task classification cases. In this work our objective is not only to develop an approach that achieves high accuracy but that also generalizes well to different cases. Henceforth, we have focused towards reducing the accuracy variation across different stress task classification cases. For our proposed approach the variation in accuracy across different stress task classification cases was reported around 9% which is approximately 6% less than the current state-of-the-art work (Ziaratnia et al., 2024).

For multi-level stress classification task we have compared our proposed approach with two state-of-the-art approaches (Ziaratnia et al., 2024), (Xu et al., 2024). In (Ziaratnia et al., 2024) authors have computed results by using 5 fold stratified cross validation approach for computing results. Using the same strategy in our case we have achieved accuracy of 0.875 with F1-score as 0.870 while in (Ziaratnia et al., 2024) computed accuracy was 0.805 with F1-score as 0.803. Clearly, we have achieved an improvement of 7% in accuracy as compared to (Ziaratnia et al., 2024). In (Xu et al., 2024), authors used 10-fold cross-validation for the categorization of low stress (T2) vs. high stress (T3) and achieved an accuracy of 83.83% . Following the same strategy, for our proposed architecture we have achieved comparable accuracy of 83.50%.

5 CONCLUSION AND FUTURE SCOPE

Stress assessment is crucial for applications such as driver condition monitoring, workplace productivity analysis, and tailored healthcare. The development of precise and economical stress task and level classification techniques is urgently needed. Therefore, in this work we have presented a deep tabular non-contact stress measurement technique. The suggested method concentrates on extracting the physiological parameters from the rPPG signal and improving ROI selection while dynamically moving the face inside the video frame. The suggested architecture per-

formed better than earlier approaches on the UBFC-Phy dataset in both sets of experiments. Future research may examine the possibilities of combining speech and eye gaze data with rPPG signals to assess stress.

REFERENCES

Arik, S. Ö. and Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687.

Barazi, N., Polidovitch, N., Debi, R., Yakobov, S., Lakin, R., and Backx, P. H. (2021). Dissecting the roles of the autonomic nervous system and physical activity on circadian heart rate fluctuations in mice. *Frontiers in physiology*, 12:692247.

Casado, C. Á., Cañellas, M. L., and López, M. B. (2023). Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing*, 14(4):3305–3316.

Choi, C.-H., Kim, J., Hyun, J., Kim, Y., and Moon, B. Face detection using haar cascade classifiers based on vertical component calibration.

Das, M., Bhuyan, M. K., and Sharma, L. N. (2023). Time–frequency learning framework for rppg signal estimation using scalogram-based feature map of facial video data. *IEEE Transactions on Instrumentation and Measurement*, 72:1–10.

Dolmans, T. C., Poel, M., van't Klooster, J.-W. J., and Veldkamp, B. P. (2021). Perceived mental workload classification using intermediate fusion multimodal deep learning. *Frontiers in human neuroscience*, 14:609096.

Fink, G. (2010). Stress: Definition and history. *Stress science: neuroendocrinology*, 3(9):3–14.

Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S. C., Wang, X., and Li, H. (2019). Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648.

Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., and Tsiknakis, M. (2019). Review on psychological stress detection using biosignals. *IEEE transactions on affective computing*, 13(1):440–460.

Imambi, S., Prakash, K. B., and Kanagachidambaresan, G. (2021). Pytorch. *Programming with TensorFlow: solution for edge computing applications*, pages 87–104.

Kim, N. H., Yu, S.-G., Kim, S.-E., and Lee, E. C. (2021). Non-contact oxygen saturation measurement using ycgcr color space with an rgb camera. *Sensors*, 21(18):6120.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. (2019). Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Mordvintsev, A. and Abid, K. (2017). Opencv-python tutorials documentation.

Ntalampiras, S. (2023). Model ensemble for predicting heart and respiration rate from speech. *IEEE Internet Computing*, 27(3):15–20.

Pan, Y., Shang, Y., Liu, T., Shao, Z., Guo, G., Ding, H., and Hu, Q. (2024). Spatial–temporal attention network for depression recognition from facial videos. *Expert Systems with Applications*, 237:121410.

Panigrahi, A. and Sharma, H. (2022). Non-contact hr extraction from different color spaces using rgb camera. In *National Conference on Communications (NCC)*, pages 332–337.

Praveen, R. G., Granger, E., and Cardinal, P. (2021). Cross attentional audio-visual fusion for dimensional emotion recognition. In *16th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–8.

Sabour, R. M., Benezeth, Y., De Oliveira, P., Chappe, J., and Yang, F. (2021). Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14(1):622–636.

Selesnick, I. W. and Burrus, C. S. (1998). Generalized digital butterworth filter design. *IEEE Transactions on signal processing*, 46(6):1688–1694.

Speth, J., Vance, N., Sporrer, B., Niu, L., Flynn, P., and Czajka, A. (2024). Mspm: A multi-site physiological monitoring dataset for remote pulse, respiration, and blood pressure estimation. *arXiv preprint arXiv:2402.02224*.

Wang, W., Den Brinker, A. C., Stuijk, S., and De Haan, G. (2016). Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491.

Xu, J., Song, C., Yue, Z., and Ding, S. (2024). Facial video-based non-contact stress recognition utilizing multi-task learning with peak attention. *IEEE Journal of Biomedical and Health Informatics*.

Yu, H., Vaessen, T., Myin-Germeys, I., and Sano, A. (2021). Modality fusion network and personalized attention in momentary stress detection in the wild. In *9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8.

Zhang, X., Wei, X., Zhou, Z., Zhao, Q., Zhang, S., Yang, Y., Li, R., and Hu, B. (2023). Dynamic alignment and fusion of multimodal physiological patterns for stress recognition. *IEEE Transactions on Affective Computing*.

Ziaratnia, S., Laohakangvalvit, T., Sugaya, M., and Sripian, P. (2024). Multimodal deep learning for remote stress estimation using cct-lstm. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8336–8344.