Supplementary Material

Kaveti Pavan Kumar
Biomedical Engineering Department
Indian Institute of Technology Hyderabad
bm23resch01001@iith.ac.in

Nagarajan Ganapathy
Biomedical Engineering Department
Indian Institute of Technology Hyderabad
gnagarajan@bme.iith.ac.in

1. Physiological Signal Visualization

To qualitatively analyze the characteristics of the collected physiological signals, we present representative 10-second segments from two randomly selected subjects under normal and stress conditions. The signals include ECG, thoracic respiration (RSP-Tho), and abdominal respiration (RSP-Abd), illustrating modality-specific variations across different stress states.

The ECG signals exhibit clear variations in waveform morphology and inter-beat intervals between normal and stress conditions, indicating changes in cardiac dynamics. Under stress, the ECG shows increased frequency and reduced variability, reflecting heightened sympathetic activity.

Respiration signals demonstrate distinct behavioral changes across conditions. Both thoracic and abdominal respiration exhibit more rapid and irregular breathing patterns under stress, with noticeable variations in amplitude and rhythm. Additionally, differences between thoracic and abdominal respiration highlight complementary physiological responses, emphasizing the importance of multimodal analysis.

These visual patterns confirm that the collected physiological signals effectively capture stress-induced variations, supporting their suitability for multimodal driver stress detection.

2. Architectural Details

In this work, multiple temporal deep learning architectures are employed to comprehensively evaluate the proposed Cross-SECA framework under real-world physiological signal conditions. The dataset used in this study is collected in real-time using wearable sensors and represents a private dataset with inherent variability across subjects and driving scenarios. Due to these constraints, all selected architectures are carefully adapted to align with the characteristics of the dataset, including signal length, sampling rates, and sample size. While preserving the fundamental design principles and novelty of each architecture, the network depth, number of filters, and layer configurations are adjusted to ensure stable training and effective generalization. This design strategy enables a fair and consistent evaluation of different temporal modeling paradigms within the same experimental setting.

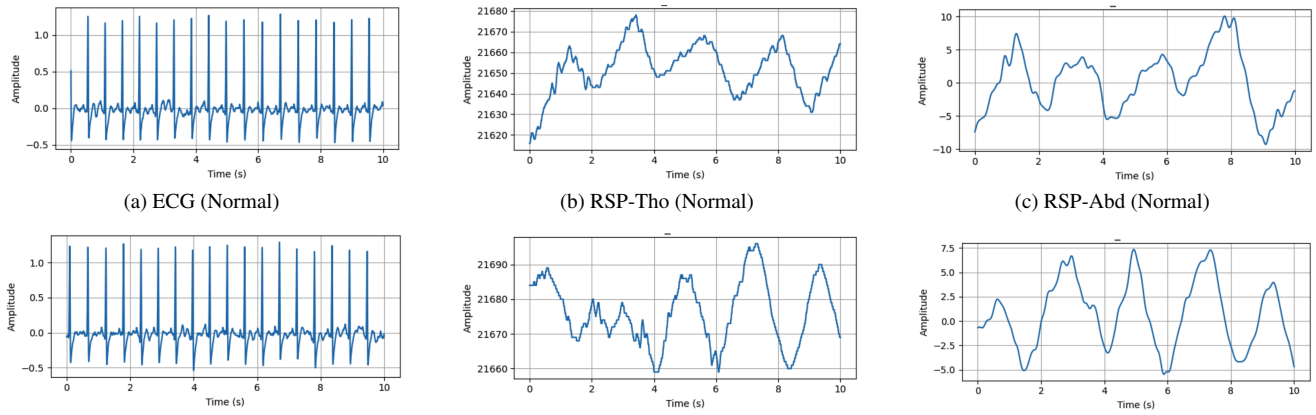
2.1. 1D Convolutional Neural Network (1D CNN)

The 1D CNN architecture is designed to capture hierarchical temporal patterns from multimodal physiological signals, specifically ECG and respiration (RSP). Separate modality-specific branches are employed for feature extraction, followed by feature-level fusion and classification.

Each modality branch consists of four sequential convolutional blocks. The first block applies a 1D convolution with 16 filters and a kernel size of 9, followed by batch normalization and average pooling with a pool size of 4. The second block increases the number of filters to 32, followed by batch normalization, average pooling, and a dropout layer with a rate of 0.3 to mitigate overfitting. The third block further increases the filters to 64, followed by batch normalization, average pooling, and dropout with a rate of 0.3. The final convolutional block uses 128 filters with a kernel size of 9, followed by batch normalization and average pooling. The resulting feature maps are flattened to obtain compact modality-specific feature representations.

The flattened features from ECG and RSP branches are concatenated to form a unified multimodal representation. This fused representation is passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units, each using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers to improve generalization. The final output layer consists of two neurons with sigmoid activation for binary classification.

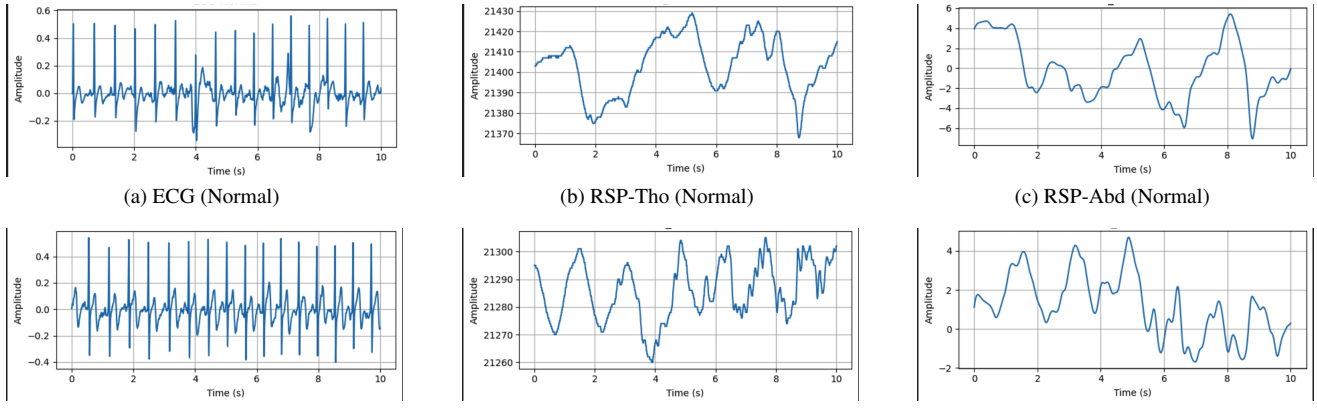
All convolutional layers use ReLU activation, and the model is trained using the RMSProp optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss.



(a) ECG (Normal) (b) RSP-Tho (Normal) (c) RSP-Abd (Normal)

(d) ECG (Stress) (e) RSP-Tho (Stress) (f) RSP-Abd (Stress)

Figure 1. Representative 10-second physiological signals from Subject 1 under normal (top row) and stress (bottom row) conditions across ECG, thoracic respiration, and abdominal respiration modalities.



(a) ECG (Normal) (b) RSP-Tho (Normal) (c) RSP-Abd (Normal)

(d) ECG (Stress) (e) RSP-Tho (Stress) (f) RSP-Abd (Stress)

Figure 2. Representative 10-second physiological signals from Subject 2 under normal (top row) and stress (bottom row) conditions across ECG, thoracic respiration, and abdominal respiration modalities.

2.2. InceptionTime

The InceptionTime architecture is employed to capture multi-scale temporal features from physiological signals by leveraging parallel convolutional operations with varying receptive fields. Similar to the 1D CNN design, separate modality-specific branches are used for ECG and respiration (RSP) signals.

Each branch consists of a sequence of Inception blocks. An Inception block comprises three parallel 1D convolutional layers with kernel sizes of 9, 19, and 39, respectively, each using the same number of filters. The outputs of these parallel convolutions are concatenated along the channel dimension, followed by batch normalization to stabilize training and improve feature representation.

In the first stage, the Inception block uses 32 filters per convolution branch, followed by average pooling with a pool size of 2. The second stage increases the number of filters to 64, followed by average pooling. The third stage further increases the filters to 128, again followed by average pooling. This progressive increase in filters enables the network to learn increasingly complex temporal patterns at multiple scales.

The final feature maps from each modality branch are flattened to obtain compact feature representations. These modality-specific features are concatenated to form a unified multimodal representation, which is then passed to a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers to reduce overfitting. The final output layer consists of two neurons with sigmoid activation for binary classification.

All convolutional layers use ReLU activation, and the model is trained using the RMSProp optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss.

2.3. Temporal Convolutional Network (TCN)

The Temporal Convolutional Network (TCN) architecture is utilized to model long-range temporal dependencies in physiological signals through dilated convolutions. Separate modality-specific branches are employed for ECG and respiration (RSP) signals, enabling independent temporal feature extraction.

Each modality branch consists of a sequence of four temporal convolutional blocks with increasing dilation rates. Each block applies a 1D convolution with a kernel size of 7, followed by batch normalization and dropout with a rate of 0.3 to improve generalization. The dilation rates are progressively increased as 1, 2, 4, and 8 across the four blocks, allowing the network to capture temporal dependencies over multiple receptive fields without increasing computational complexity.

The number of filters is increased across layers from 16 to 32, 64, and 128, enabling hierarchical feature learning from low-level to high-level temporal patterns. After the final convolutional block, average pooling with a pool size of 2 is applied to reduce the temporal dimension, followed by flattening to obtain compact modality-specific feature representations.

The flattened features from both modalities are concatenated to form a unified multimodal representation. This fused representation is passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers. The final output layer consists of two neurons with sigmoid activation for binary classification.

All convolutional layers use ReLU activation, and the model is trained using the RMSProp optimizer with a learning rate of 1×10^{-4} and binary cross-entropy loss.

2.4. MobileNet-1D

The MobileNet-1D architecture is adopted to efficiently model temporal patterns in physiological signals using depthwise separable convolutions, which significantly reduce computational complexity while maintaining representational capacity. Similar to other architectures, separate modality-specific branches are used for ECG and respiration (RSP) signals.

Each branch consists of a sequence of depthwise separable convolutional blocks. Each block is composed of a depthwise 1D convolution with a kernel size of 5, followed by batch normalization, and a pointwise 1D convolution with a kernel size of 1 for channel mixing, again followed by batch normalization. All convolutional layers use ReLU activation.

The network progressively increases the number of filters across four stages with 16, 32, 64, and 128 filters, enabling hierarchical feature extraction from low-level to high-level temporal representations. Unlike conventional CNN architectures, no explicit pooling layers are used, allowing the network to preserve temporal resolution while learning compact representations through convolutional operations. The final feature maps are flattened to obtain modality-specific feature vectors.

The flattened features from ECG and RSP branches are concatenated to form a unified multimodal representation. This fused representation is passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers to improve generalization. The final output layer consists of two neurons with sigmoid activation for binary classification.

2.5. ConvMixer-1D

The ConvMixer-1D architecture is employed to capture temporal dependencies in physiological signals through a combination of patch embedding and repeated mixing operations. Separate modality-specific branches are used for ECG and respiration (RSP) signals to enable independent feature extraction.

Each branch begins with a patch embedding layer implemented using a 1D convolution with 16 filters, a kernel size of 5, stride 1, and ReLU activation, followed by batch normalization. This layer transforms the raw input signals into a higher-dimensional feature space suitable for subsequent processing.

Following the embedding stage, a series of ConvMixer blocks are applied. Each ConvMixer block consists of a depthwise 1D convolution with a kernel size of 5 for temporal (spatial) mixing, followed by batch normalization, and a pointwise 1D convolution with a kernel size of 1 for channel mixing, again followed by batch normalization. Dropout with a rate of 0.3 is applied after each ConvMixer block to improve generalization. The number of filters is progressively increased across blocks to 32, 64, and 128, enabling hierarchical feature learning.

Unlike conventional architectures, no explicit pooling layers are used, allowing the network to preserve temporal resolution while learning discriminative features through convolutional mixing operations. The final feature maps are flattened to obtain modality-specific feature representations.

The flattened features from ECG and RSP branches are concatenated to form a unified multimodal representation. This fused representation is passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers. The final output layer consists of two neurons with sigmoid activation for binary classification.

2.6. 1D ResNet

The 1D ResNet architecture is employed to learn deep hierarchical temporal representations from physiological signals using residual learning. Separate modality-specific branches are designed for ECG and respiration (RSP) signals to enable independent feature extraction.

Each branch consists of a sequence of four residual blocks. Each residual block comprises two consecutive 1D convolutional layers with a kernel size of 9. The first convolution uses ReLU activation followed by batch normalization, while the second convolution is followed by batch normalization without activation. A residual (shortcut) connection is added between the input and output of the block to facilitate gradient flow and stabilize deep network training. When the number of filters changes across blocks, a 1×1 convolution is applied to the shortcut connection to match the dimensionality.

The number of filters is progressively increased across the residual blocks as 16, 32, 64, and 128, enabling the network to learn increasingly abstract temporal features. After the final residual block, the feature maps are flattened to obtain modality-specific feature representations.

The flattened features from ECG and RSP branches are concatenated to form a unified multimodal representation. This fused representation is passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers to improve generalization. The final output layer consists of two neurons with sigmoid activation for binary classification.

2.7. DenseNet

The DenseNet architecture is employed to enhance feature propagation and reuse through dense connectivity, enabling efficient learning of temporal representations from physiological signals. Separate modality-specific branches are designed for ECG and respiration (RSP) signals.

Each branch begins with an initial 1D convolutional layer with 32 filters and a kernel size of 7, followed by ReLU activation. This is followed by a series of dense blocks, where each dense block consists of multiple convolutional layers with dense connectivity. In each dense block, every layer receives input from all preceding layers via feature concatenation, promoting feature reuse and improving gradient flow.

Each dense block contains four layers, with a growth rate of 16. Each layer within the dense block applies batch normalization, ReLU activation, and a 1D convolution with a kernel size of 3. The output of each layer is concatenated with the input feature maps, progressively increasing the number of channels.

Between dense blocks, a transition layer is applied to reduce feature dimensionality. The transition layer consists of batch normalization, ReLU activation, a 1D convolution with a kernel size of 1, and average pooling with a pool size of 2. A reduction factor of 0.5 is used to compress the number of feature channels.

After the final dense block, the feature maps are flattened to obtain modality-specific feature representations. The features from ECG and RSP branches are concatenated to form a unified multimodal representation.

The fused features are passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers to improve generalization. The final output layer consists of two neurons with sigmoid activation for binary classification.

2.8. Transformer

The Transformer architecture is employed to model long-range temporal dependencies in physiological signals using self-attention mechanisms. Separate modality-specific branches are designed for ECG and respiration (RSP) signals.

Each branch begins with a linear embedding layer that projects the input signal into a higher-dimensional feature space with an embedding dimension of 64. To retain temporal ordering information, sinusoidal positional encoding is added to the embedded features.

The encoded representations are then passed through a stack of Transformer encoder blocks. Each block consists of a multi-head self-attention mechanism followed by a position-wise feed-forward network. The multi-head attention uses 4 attention heads with a key dimension of 16, enabling the model to capture diverse temporal dependencies across different subspaces. Residual connections and layer normalization are applied after both the attention and feed-forward sublayers to stabilize training and improve convergence. Dropout with a rate of 0.3 is applied to both attention outputs and feed-forward layers to reduce overfitting.

Each feed-forward network consists of two fully connected layers, where the first layer expands the feature dimension to 128 using ReLU activation, followed by a projection back to the original embedding dimension.

After passing through two Transformer encoder blocks, the feature maps are flattened to obtain modality-specific representations. The features from ECG and RSP branches are concatenated to form a unified multimodal representation.

The fused features are passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers. The final output layer consists of two neurons with sigmoid ac-

tivation for binary classification.

2.9. ConvTransformer

The ConvTransformer architecture combines convolutional neural networks with Transformer-based self-attention to capture both local temporal patterns and long-range dependencies in physiological signals. Separate modality-specific branches are designed for ECG and respiration (RSP) signals.

Each branch begins with convolutional feature extraction to capture local temporal structures. The first convolutional layer uses 32 filters with a kernel size of 9 and ReLU activation, followed by batch normalization and average pooling with a pool size of 4. This is followed by a second convolutional layer with an embedding dimension of 64, kernel size 9, and batch normalization, which transforms the features into a representation suitable for attention-based modeling.

The extracted features are then processed using Transformer encoder blocks. Each Transformer block consists of a multi-head self-attention mechanism with 4 heads and a key dimension of 16, followed by a position-wise feed-forward network with a hidden dimension of 128. Residual connections and layer normalization are applied after both the attention and feed-forward sublayers to stabilize training. Dropout with a rate of 0.3 is applied within the Transformer blocks to improve generalization. Two such Transformer blocks are stacked to capture global temporal dependencies.

The output feature maps from each modality branch are flattened to obtain compact representations. The features from ECG and RSP branches are concatenated to form a unified multimodal representation.

The fused features are passed through a fully connected classification network consisting of dense layers with 128, 64, 32, and 16 units using ReLU activation. Dropout with a rate of 0.3 is applied after the first two dense layers. The final output layer consists of two neurons with sigmoid activation for binary classification.

3. Implementation Details

All models are implemented using the TensorFlow deep learning framework and trained in a subject-independent manner using Leave-One-Subject-Out Cross-Validation (LOSOCV). For each fold, data from one subject is used for testing while the remaining subjects are used for training. The models are trained for 30 epochs with a batch size of 18 using binary cross-entropy loss. The RMSProp optimizer with a learning rate of 1×10^{-4} is used for convolution-based architectures, while the Adam optimizer with the same learning rate is employed for Transformer-based models. Dropout with a rate of 0.3 is applied across networks to mitigate overfitting. Model performance is evaluated using accuracy, F1-score, precision, and recall metrics, computed by averaging results across all LOSOCV folds. All computations were executed using Python programming language on a Dell Precision 7960 workstation equipped with an Nvidia RTX A5000 24GB GPU.