# AT2GRU: A Human Emotion Recognition Model with Mitigated Device Heterogeneity

Pritam Khan, Priyesh Ranjan, and Sudhir Kumar, *Senior Member, IEEE*

**Abstract**—Device heterogeneity can cause a detrimental impact on the classification of healthcare data. In this work, we propose the Maximum Difference-based Heterogeneity Mitigation (MDHM) method to address device heterogeneity. Mitigating heterogeneity increases the reliability of using multiple devices from different manufacturers for measuring a particular physiological signal. Further, we propose an attention-based bilevel GRU (Gated Recurrent Unit) model, abbreviated as AT2GRU, to classify multi-modal healthcare time-series data for human emotion recognition. The physiological signals of Electroencephalogram (EEG) and Electrocardiogram (ECG) for twenty-three persons are leveraged from the DREAMER dataset for emotion recognition. Also, from the DEAP dataset, the biosignals namely EEG, Galvanic Skin Response (GSR), Respiration Amplitude (RA), Skin Temperature (ST), Blood Volume (BV), Electromyogram (EMG) and Electrooculogram (EOG) of thirty-two persons are used for emotion recognition. The EEG and the other biosignals are denoised by the wavelet filters for enhancing the model's classification accuracy. A multi-class classification is carried out considering valence, arousal, and dominance for each person in the datasets. The classification accuracy is validated against the self-assessment obtained from the respective person after watching a movie/video. The proposed AT2GRU model surpasses the other sequential models namely Long Short Term Memory (LSTM) and GRU in performance.

**Index Terms**—Attention, electrocardiogram, electroencephalogram, emotion, GRU, healthcare.

## 1 Introduction

Human emotion results from the reaction in the nervous system of the body and it depends on the circumstances that a person encounters. Prediction of human responses to stimuli or understanding a person's mental well being are essential in today's world keeping in mind the adverse effects of depression. Accurate prediction of human emotion recognition is a difficult task involving complex and robust predictive algorithms. Involvement of artificial intelligence in human emotion prediction has expedited the research on human psychology [1], [2], [3], [4]. The features extracted from the biosignals, image, video, audio, and text using various machine learning algorithms can be analyzed for the prediction of human emotion [1], [2], [3], [5], [6], [7]. Among the various biosignals, the EEG (Electroencephalogram) signals are used by many works for recognizing the human emotion [1], [2], [3], [8], [9], [10]. However, to improve the classification accuracy, other biosignals like ECG (Electrocardiogram), EMG (Electromyogram), skin response, and respiration are also leveraged and features are extracted from the same [2], [9], [11]. The EEG and the other biosignals' features are generally analyzed using Recurrent Neural Network (RNN) and its variants like LSTM (Long Short Term Memory)/GRU (Gated Recurrent Unit) as this category of neural networks specialize in handling time-series data. Notably, ordinary RNNs have vanishing and exploding gradient problems whereas LSTM or GRU cell is free from them owing to the additive effects instead of multiplicative gradients during back-propagation [12]. This prohibits RNNs from learning long term dependencies in data which LSTM/GRU cells can learn thereby helping in classification of data with time dependencies. However, the GRU is free from the complexity of having a memory unit unlike LSTM and is also fast as compared to LSTM. A GRU directly exposes the hidden content inside without having to keep anything in memory [13]. Therefore, in this work, we use GRU to analyze the time-series EEG and other physiological signals. Although LSTM/GRU can be used for emotion recognition task leveraging the physiological signals, the classification performance can get improved by paying more attention to the significant signals for a particular emotion. The concept of attention mechanism is motivated from the property of human perception. Humans focus on a certain portion of the visual space to capture the necessary information instead of assessing the entire scene all at once [9]. This focusing helps in the decision making. For example, each EEG frequency band extracted from the EEG signal is dominant during a particular emotion [1], [2], [14], [15]. Hence, paying more attention to the dominant EEG frequency band enhances the classification accuracy of the corresponding elicited emotion. Giving equal weightage to all the signals irrespective of their contribution for emotion recognition limits the effectiveness of classification model. Attention model enables the selection of features stressing on the important ones based on the desired output. An attention-based GRU architecture is discussed by [16] for answering visual and textual questions. Therefore, in this work we design an attention-based architecture using GRU so as to prioritize the biosignals based on their significance in determining a particular emotion. This significance is learnt by the model during training on back-propagating. The attention-based bilevel gated recurrent unit (AT2GRU)

- *P. Khan, P. Ranjan and S. Kumar are with the Department of Electrical Engineering, Indian Institute of Technology Patna, Bihar, 801106 India. Email: {pritam_1921ee05, priyesh.ee17, sudhir}@iitp.ac.in.*

*This work acknowledges the support rendered by the Early Career Research (ECR) award scheme project "Cyber-Physical Systems for M-Health" (ECR/2016/001532) (duration 2017-2020), under Science and Engineering Research Board (SERB), Govt. of India.*

model proposed in our work further improves the classification performance metrics of a GRU.

However, the researchers and the medical professionals need to be aware of the effects of the device heterogeneity while acquiring the physiological signals from different devices [17]. Device heterogeneity can result from sensor biases (due to poor calibration, limited granularity and range, poor repeatability, or accidental device dropping), sampling rate heterogeneity (varied default sampling frequencies of sensors across devices from different manufacturers), or sampling rate instability (timestamp delay between sensor measurements and instantaneous input/output load on a device) [17]. This can slow down the training process and/or lead to inconsistency in the classification results thereby impacting on the reliability and versatility of the classification models [17], [18]. As per the report of U.S. Food and Drug Administration of 2011 [19], heterogeneity and complexity of the devices from different manufacturers caused adverse effects. Almost 60 percent of adverse events are reported from cardiovascular, in-vitro diagnostic, and other medical devices [19]. Therefore, we propose the MDHM method to mitigate the device heterogeneity issue thereby facilitating the use of devices from different manufacturers for measuring a physiological signal. The issue of device heterogeneity is hardly addressed for medical devices and the proposed MDHM method will be of immense help for handling the medical data from multiple devices yielding heterogeneous results.

Also, we use multiple wavelet filters for denoising the raw physiological data. The classification accuracy of denoised data from each filter differs from each other although they are supposed to have similar performance. Therefore, we consider the filter yielding the highest accuracy in each case. Our results are validated on two standard human emotion recognition databases using the proposed AT2GRU model.

### 1.1 Related Work

The emotions can be classified using either a categorical approach (into discrete classes like joy, anger, happy, sad, etc.) or a dimensional approach (into numerical values over a number of emotion dimensions like valence and arousal) [20]. The word and emoticon lexicon approach, keyword spotting approach, and machine learning approach are commonly used for emotion recognition from text [7]. Most of the text-based emotion classification research works consider word as an important feature in building the classification models [21]. Utilizing the Synesketch software, [7] investigates the similarity between human rating and corresponding computer generated emotional rating of sentences through correlation and cosine similarity. In a meta-analytic investigation of autonomic features of the emotion categories, [22] discusses that each emotion category results in certain coordinated changes of facial muscles, activity of autonomic nervous system (like heart rate, respiration, perspiration, etc.), and affective quality of experience. There are multiple works in the literature for recognition of human emotion of subjects based on EEG signals. However, multimodal physiological signals are also used by many works for improving the classification accuracy.

Most of the works are based on the well-known datasets like DREAMER [1], DEAP [2], MPED [9], MAHNOB-HCI [11]. We validate our proposed work using the DREAMER and the DEAP datasets that are obtained through similar experimental procedures and classified in similar way. The state-of-the-art works on emotion recognition leveraging these two datasets are tabulated in Table 1. The accuracies being in terms of valence/arousal/dominance are different. Therefore, considering all the dimensional accuracies of the works in Table 1, we decide the threshold accuracy as 78 percent to categorize the accuracies into medium or high. Valence decides whether the emotion is positive/high (happiness) or negative/low (sadness). High to low values of arousal indicate the corresponding emotion ranging from excitement to boredom. The excitement/high arousal can be a case of joy or anger, while sadness or boredom corresponds to low arousal. A low dominance value indicates submissiveness like fear, whereas a high dominance denotes a feeling in control, like admiration.

TABLE 1: Overview of the literature using DREAMER/DEAP datasets for human emotion recognition

| Classification Method | Dataset | Number of Classes | Accuracy |
|---|---|---|---|
| RBF-based SVM [1] | DREAMER | Binary (2) | Medium |
| Gaussian naive Bayes [2] | DEAP | Binary (2) | Medium |
| GraphSLDA [8] | DREAMER | Binary (2) | Medium |
| GSC Correlation Analysis [8] | DREAMER | Binary (2) | Medium |
| Dynamical Graph CNN [8] | DREAMER | Binary (2) | High |
| GCB-net [3] | DREAMER | Binary (2) | High |
| MMResLSTM [23] | DEAP | Binary (2) | High |
| H-ATT-BGRU [24] | DEAP | Binary (2) | Medium |
| Extreme Learning Machine [6] | DREAMER/DEAP | Binary (2) | High/High |
| Deep forest [10] | DREAMER/DEAP | Binary (2) | High/High |
| BioCNN [25] | DEAP | Binary (2) | High |
| **AT2GRU [Proposed]** | DREAMER/DEAP | **Multiple (5/9)** | **High/High** |

In [1], the subjects are asked to grade between 1 to 5 the values of valence, arousal, and dominance after watching each movie-clip. This grading is tallied with the corresponding values obtained by using the physiological signal features and hence, the accuracies are obtained by comparison. The classification obtained from the signals is binary (low/high) and not multi-class thereby being a limitation for the process used for classification. A similar experiment is carried out in [2] except that the grading done by the subjects is in the range 1 to 9. Here also, the prediction is performed on a binary scale (low/high) based on a threshold. EEG-based emotion classification using the novel Dynamical Graph Convolutional Neural Networks (DGCNN) is discussed in [8] using the DREAMER database. The authors achieve higher classification accuracies for valence, arousal, and dominance using DGCNN as compared to the state-of-the-art Support Vector Machine (SVM), Graph regularized Sparse Linear Discriminant Analysis (Graph-SLDA), and Group Sparse Canonical Correlation Analysis (GSCCA). Only EEG data of the DREAMER dataset is used for emotion recognition without leveraging the available ECG data of the corresponding subjects. Further, the accuracies are based on a binary classification (low/high) of the valence, arousal, and dominance. As an improvement on [8], a Graph Convolutional Broad Network (GCB-net) is proposed in [3] and higher classification accuracies are obtained using the same datasets. However, [3] again goes

for binary classification, that is, low/high on the DREAMER dataset thereby keeping the fuzziness in the results. A multi-modal emotion recognition dataset is proposed by [9] and classification is carried out using Attention-LSTM (A-LSTM). [9] uses the features of ECG, EEG, Galvanic Skin Response (GSR), and respiration for emotion recognition and achieves the best results using A-LSTM as compared to LSTM, K-Nearest Neighbour (KNN), and SVM. The time-complexity of A-LSTM is high as compared to the other classifiers. Additionally, in order to develop the attention model, the architecture of a CNN is used for each of the three masks into the LSTM thereby increasing the overall model complexity.

A Multi-modal Residual LSTM (MMResLSTM) model is proposed in [23] using the DEAP database for emotion recognition with binary classification output. High accuracy is obtained using the residual model of LSTM by tuning the hyperparameters of the model without separately extracting the features from the data. An attention-based hierarchical bidirectional GRU model is proposed by [24] for emotion recognition leveraging DEAP database, and the performance is compared with that of LSTM, CNN, and SVM. Although [24] achieves low accuracies in their experiment, the attention-based bidirectional GRU performs better than the other models. In [6], many datasets including DREAMER and DEAP are leveraged for multi-modal emotion recognition using Extreme Learning Machines (ELM) but the accuracies obtained using the DREAMER dataset are moderate and surpassed by an earlier work [8]. A deep forest model called multi-grained cascade model is used for emotion recognition using the DEAP and the DREAMER databases in [10] where the accuracies improve as compared to the other existing works. The classification carried out is based on the pre-processed data used to construct two dimensional frame sequences. However, any other pre-processing method is not verified to yield the same performance which is subjected to variation owing to the heterogeneity in the processed data. A hardware called BioCNN is discussed in [25] for emotion recognition besides using the datasets of DEAP, DREAMER, and FEXD. Although the novelty lies in framing a hardware model, the classification accuracies obtained remain limited as compared to many existing works.

Another issue that remains hardly addressed during human emotion prediction is the heterogeneity in the pre-processed data. In the literature, few works address the device heterogeneity for a consistency in the results from multiple data sources. In [17], eight smartphones and four smart watches are used for Human Activity Recognition (HAR); however, the performance gets significantly affected by sensor heterogeneity. The effect of device heterogeneity is significant in the feature extraction stage [17]. Interpolation and clustering techniques are used by [17] for device heterogeneity mitigation thereby improving the classification performance. An adaptive quantization model is discussed by [18] in a federated learning environment to mitigate device heterogeneity and improve quality (average test accuracy) and fairness (cosine similarity). [26] demonstrates a clinical decision support system capable of handling multiple heterogeneous healthcare data sources. [27] addresses the device heterogeneity using zero-mean and unity-mean features of Wi-Fi (wireless fidelity) received signal strength in a localization work. In [28], the Nelder-Mead (NM) minimization technique is discussed which can also be used for device heterogeneity mitigation.

In summary, device heterogeneity remains least addressed which can lead to change in the classification accuracies. Additionally, all the state-of-the-art works on emotion recognition, perform a binary (low/high) classification of results, while the subjects who participate in the test, generally give a multi-class rating. Therefore, the device heterogeneity is mitigated in this work so that the classification methods based on data acquired from multiple devices yield consistent results. Also, we perform a multi-class classification of human emotion based on the DREAMER and DEAP datasets with improved classification accuracies as compared to many existing works using the proposed AT2GRU model.

### 1.2 Contributions

The major contributions of our work are enlisted as follows:

1) The proposed MDHM method mitigates the device heterogeneity to increase the reliability of using the devices from multiple manufacturers for measuring a particular physiological parameter. This helps to attain a consistent classification performance irrespective of the device used.
2) We leverage the multi-modal physiological data from different datasets for emotion recognition using the proposed AT2GRU model that outperforms the common RNN models namely LSTM and GRU in performance. The proposed model pays weighted attention to the required physiological signals dominant for any particular emotion.
3) We obtain a multi-class classification of valence, arousal, and dominance for emotion recognition unlike the state-of-the-art works that perform a binary classification (low/high).

The rest of the paper is organized as follows. Section 2 discusses the heterogeneity mitigation among devices. The classification of human emotion using the proposed AT2GRU based on the multi-modal physiological signals is discussed in Section 3. In Section 4, we discuss the experiments and results highlighting the improvement in performance on using the AT2GRU model. Finally, Section 5 concludes the paper.

## 2 Mitigation of Device Heterogeneity for Consistent Classification Accuracy

In this work, we classify the human emotion based on multiple physiological signals. As a result of device heterogeneity, multiple devices measuring the same physiological signal for a person can yield different results thereby causing inconsistency in the classification accuracy. Therefore, we mitigate the heterogeneity issue in devices measuring physiological signals using the proposed MDHM method.

### 2.1 Data Pre-processing

The raw EEG and other acquired biosignals require to be denoised for accurate human emotion recognition. The

brain waves corresponding to different frequency bands namely delta, theta, alpha, beta, gamma are used for the analysis of EEG data. Various methods are used for denoising the acquired raw physiological signals to improve the classification accuracy. We denoise the raw biosignals using the wavelet filters namely Daubechies, Symlet, and Coiflet filters as we obtain slightly different classification performance from each of them. The chosen family of filters, however, have similar properties [29]. In each case, we consider the filter result yielding the maximum accuracy and use the denoised output from the filter for emotion recognition. The EEG signals are decomposed through the wavelet filters to extract the $\delta$ band (1-3Hz), $\theta$ band (4-7Hz), $\alpha$ band (8-13Hz), $\beta$ band (14-30Hz), and $\gamma$ band (31-50Hz) before being fed to the GRU [8]. The raw EEG signals get denoised by the removal of high and low frequency noise spectrum. However, the denoised EEG signals corresponding to the same frequency band including the other denoised biosignals obtained from multiple devices can vary owing to heterogeneity. The heterogeneity mitigation in the processed data helps to achieve consistency in the classification accuracy of the classifier. Prior to heterogeneity mitigation, the denoised data from the wavelet filters are upsampled and normalized to make a point-by-point comparison among samples from multiple devices within a particular range. Owing to the different range of the acquired data from various devices, the normalization of the data is carried out for comparison. Each signal is normalized in a non-zero positive range.

## 2.2 Proposed MDHM Method

Device heterogeneity can lead to inconsistency in the data classification accuracy. Hitherto, device heterogeneity is mitigated in various works as discussed in literature [17], [27], [28]. However, the interpolation and clustering techniques for heterogeneity mitigation in [17] are designed based on available devices and the performance of a new device is not guaranteed. Also, the performance of the zero-mean and unity-mean methods discussed in [27] get limited by the number of Wi-Fi (wireless fidelity) access points (AP), that is, the heterogeneity is shown to decrease with increase in the number of APs. Therefore, in this work, we propose the MDHM method which can mitigate the device heterogeneity among multiple devices without being dependent on the number of devices. The NM method being independent of number of devices, is also used by us for heterogeneity mitigation [28]. The NM method minimizes a function of $c$ variables by comparing the function values at $(c+1)$ vertices of a general simplex and then replacing the vertex having the highest value by another point. The transformation of sample points (vertices of simplex) are carried out by reflection, expansion, contraction, or shrinking depending on the conditions as described in [28]. However, on comparison with the proposed MDHM method, we find the latter to outperform the NM method. The proposed MDHM method is applied on the denoised EEG signals corresponding to the different frequency bands and on the other denoised biosignals. Although we discuss the heterogeneity mitigation for the purpose of affective computing, the method is applicable to other cases also, where the health data is acquired using different devices.

### 2.2.1 MDHM Algorithm

We illustrate the MDHM method using one denoised sample namely, $k$th sample from $n$ medical devices measuring the same physiological signal of a person. Using the Manhattan and Euclidean distance measures, we mitigate the device heterogeneity. The Euclidean distances for the $k$th samples corresponding to all devices are calculated from a reference axis and each $k$th sample is divided by that Euclidean distance value $x_{k,\text{Euc}}$ for normalizing it within 1. Next, we calculate the difference of all the other device outputs from each device output. Then we record the maximum difference obtained between the $k$th sample of every device with respect to another device. The minimization of error is carried out considering this maximum difference value from every denoised device output. For update, the sample value is multiplied by the decreasing maximum difference at every iteration. For every difference between two samples, we consider the absolute value to avoid the effect of sign change as each sample value gets modified in subsequent iterations. The iterations continue till the optimal number of iterations $l^\star$ is reached, where $l^\star = \underset{l}{\operatorname{argmin}} \left( \Delta x_{k,ij}^{(l-1)} - \Delta x_{k,ij}^{(l)} \right) = \underset{l}{\operatorname{argmin}} \left( \left| x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)} \right| - \left| x_{k,i}^{(l)} - x_{k,j}^{(l)} \right| \right)$, that is, when the difference between the $k$th samples from $i$th and $j$th devices is the minimum between $(l-1)$th and $l$th iterations. The error which is the Manhattan distance between the samples from two devices, decreases and stabilizes at a point with increase in number of iterations, and we stop iterating at the optimal $l^\star$th iteration so that the unnecessary computations are avoided. The proposed MDHM method, described by the Algorithm 1, is executed for all the samples corresponding to each device.

**Algorithm 1** Proposed MDHM technique

**Data:** Pre-processed upsampled data from $n$ devices
**Result:** Heterogeneity mitigated data

1) Collect the denoised upsampled data from $n$ devices. Let the outputs from $n$ devices be $x_{k,1}, x_{k,2}, \ldots, x_{k,n}$ for the $k$th sample.
2) Find the Euclidean distance of the $k$th samples corresponding to $n$ devices from the reference axis (say, x-axis for 1 dimensional signals). $x_{k,\text{Euc}} = \sqrt{\sum_{i=1}^{n} x_{k,i}^2} \;\; \forall \;\; i \in \{1, 2, \ldots, n\}$ where $x_{k,i}$ is the $k$th sample from $i$th device.
3) Calculate the Manhattan distance among the $k$th samples of all devices, and note the highest value corresponding to each device. For the $i$th device, the maximum difference value of the sample from that of any other device is $\Delta x_{k,i}^{(l)} = \max(\left| x_{k,i}^{(l)} - x_{k,j}^{(l)} \right|), i \neq j, \forall \;\; i, j \in \{1, 2, \ldots, n\}$ at the $l$th iteration.
4) Divide the $k$th sample of each device by the Euclidean distance measure $x_{k,\text{Euc}}$ and multiply by the maximum difference value from another device, that is, $\Delta x_{k,i}^{(l)}$.
5) At the $l$th iteration, we have the updated $k$th sample of $i$th device as $x_{k,i}^{(l)} = (x_{k,i}^{(l-1)} / x_{k,\text{Euc}}^{(l-1)}) \times \Delta x_{k,i}^{(l-1)}$.
6) Iterate from Step 2 until $l^\star$th iterationis reached.
7) Repeat Step 2 to 6 for each of the samples from all the devices.

#### 2.2.2 Convergence Analysis of MDHM Method

Using the MDHM method, we mitigate the heterogeneity among the denoised signals from the medical devices to get a consistency in the classification accuracy. Let $x_{k,i}^{(l)}$ and $x_{k,j}^{(l)}$ be the $k$th samples from $i$th and $j$th devices respectively after the $l$th iteration. For convenience, we consider that these two device outputs have the maximum difference as compared to the other devices during $l$th and $(l-1)$th iterations. Let us assume that we are considering the update of the $i$th device output. Therefore, at the $l$th iteration, the difference between them is $\left|x_{k,i}^{(l)} - x_{k,j}^{(l)}\right|$. The maximum difference equation can be explicitly written as:

$$\left|x_{k,i}^{(l)} - x_{k,j}^{(l)}\right| = \left| \frac{x_{k,i}^{(l-1)}\left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right|}{x_{k,\text{Euc}}^{(l-1)}} - \frac{x_{k,j}^{(l-1)}\left|x_{k,j}^{(l-1)} - x_{k,i}^{(l-1)}\right|}{x_{k,\text{Euc}}^{(l-1)}} \right| \tag{1}$$

Equation 1 can be further written as:

$$\left|x_{k,i}^{(l)} - x_{k,j}^{(l)}\right| = \frac{\left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right|}{x_{k,\text{Euc}}^{(l-1)}} \left|x_{k,j}^{(l-1)} - x_{k,i}^{(l-1)}\right| \tag{2}$$

Now, the difference $\left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right|$ is less than or equal to $x_{k,\text{Euc}}^{(l-1)}$ as the latter one is additive. In case $i$ and $j$ are the only 2 devices present, then $x_{k,\text{Euc}}^{(l-1)} = \sqrt{(x_{k,i}^{(l-1)})^2 + (x_{k,j}^{(l-1)})^2}$. Now we can write:

$$\left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right| = \sqrt{(x_{k,i}^{(l-1)})^2 + (x_{k,j}^{(l-1)})^2 - 2x_{k,i}^{(l-1)}x_{k,j}^{(l-1)}} \tag{3}$$

Hence, from Equation 3 we can write:

$$\left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right| \leq \sqrt{(x_{k,i}^{(l-1)})^2 + (x_{k,j}^{(l-1)})^2} \tag{4}$$

Therefore, we have $\frac{\left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right|}{x_{k,\text{Euc}}^{(l-1)}} \leq 1$. Hence, from Equation 2, we can infer that:

$$\left|x_{k,i}^{(l)} - x_{k,j}^{(l)}\right| \leq \left|x_{k,i}^{(l-1)} - x_{k,j}^{(l-1)}\right| \tag{5}$$

which means that the difference between the samples in $l$th iteration is less than that of $(l-1)$th iteration. The iterations stop when the difference between the samples for $l$th and $(l-1)$th iterations are equal. This confirms that the heterogeneity mitigates among the samples as the number of iterations increases.

#### 2.2.3 Computational Complexity of the MDHM Method

The difference from one device output to all other devices is recorded corresponding to every sample. Therefore, in every iteration for each of the $n$ devices we have $(n-1)$ differences calculated, which means we have a total of $n(n-1)$ subtractions. In one heterogeneity mitigation step, we have one subtraction corresponding to the maximum difference, one multiplication, one division, one square root, $n$ times multiplication (for squaring to get the Euclidean distance in denominator), and $(n-1)$ times addition (while calculating Euclidean distance). Considering every single mathematical operation as one Floating-point Operation (FLOP), we have a total of $n(3n+2)$ FLOPs in one iteration for all the $n$ devices. Assuming that there are $l$ iterations, the computational complexity of the proposed MDHM algorithm for the samples at one instant is $O(3ln^2 + 2ln)$.

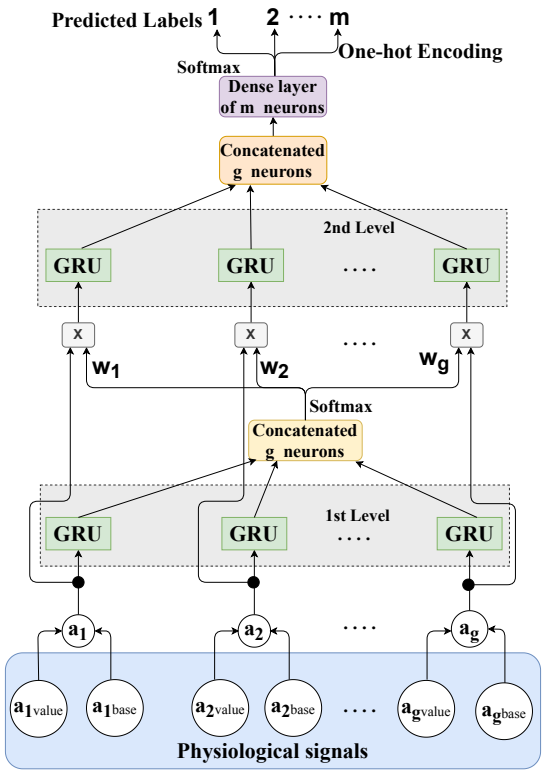



Fig. 1: Multi-modal multi-class human emotion recognition using proposed AT2GRU model

## 3 AT2GRU-based Emotion Recognition

The raw multi-modal physiological signals are denoised using the wavelet filters and then used for recognition of human emotion. The heterogeneity mitigation method is applied to make the classifier robust on the acquired signals obtained using the device from any manufacturer. Mitigation of device heterogeneity helps to bring a consistency in the classification accuracy for the data obtained from different devices. In the emotion recognition experiment, we consider each signal value and its corresponding baseline value to finally obtain a value of the valence/arousal/dominance based on which an emotion gets classified. The proposed AT2GRU model improves the classification accuracy over the traditional GRU or LSTM. The emotion recognition method being a non-invasive one, can be used by the researchers/medical professionals to study and analyze the psychological condition and nervous system of patients.

### 3.1 Proposed AT2GRU Model

Based on the different movies, various emotions namely calmness, surprise, amusement, fear, excitement, disgust, happiness, anger, and sadness are elicited each of which has a grading for valence, arousal, and dominance, say in the range of 1 to 5 or 1 to 9, based on the dataset used. The EEG and other physiological signals of the subjects participating in the experiment, are recorded corresponding to each movie and are analyzed and classified for emotion

recognition. As EEG and the other acquired biosignals comprise time-series data, therefore unlike many works leveraging the features from the images or manually acquiring the features, we use GRUs that are specialized for handling time-series data. The roles of forget gate and input gate of LSTM are taken up by the update gate of GRU [13]. GRU also has another gate called reset gate for resetting the last hidden state output if required. Both the gates of GRU are sigmoid activated while the update of the hidden-state is $\tanh$ activated. In this work, we use GRU instead of LSTM as we address the multi-modal signals in small batches thereby eliminating the presence of long time-series data. Notably, GRU works well on small-sized data while LSTM is suitable for a large dataset owing to the presence of memory-cell that makes it trade off with the complexity.

In order to enhance the classification accuracy of a model, it is important that more attention is given to the biosignal that helps in classification of a particular emotion. Therefore, in this work, we propose the AT2GRU model for improving the emotion recognition accuracy.

Figure 1 shows the proposed AT2GRU model for multi-modal human emotion recognition. The GRUs are used at two levels in the proposed attention-based model. The first level of GRUs contribute in identifying the vital features in a physiological signal and assign appropriate weightage to it for identifying a particular emotion. On the other hand, the second level of GRUs are used to classify the emotion dimension (valence/arousal/dominance) based on the attention-weighted physiological signals. At the first level, $g$ number of GRUs are used, each of which takes a biosignal $\mathbf{a}_i$ as input where $i \in \{1, 2, \ldots, g\}$. Each signal is denoised prior to feature extraction and classification through the AT2GRU model. The base values of each signal with suffix 'base' are subtracted from the actual corresponding values with suffix 'value' to generate the particular signal vector as shown in Figure 1. The base values are subtracted as they contribute to the neutral emotions whereas our requirement is to predict the emotion corresponding to each movie clipping. The output of the final hidden states of the first level GRUs are concatenated thereafter. A softmax activation is applied at the concatenation layer for classification into $g$ classes, each of which corresponds to each input signal. The softmax-activated concatenation layer yields the trainable attention weights as outputs which get updated every time during back-propagation while minimizing the prediction error during training.

At the second level, there are again $g$ GRUs each of which receives a weighted input as the product of the original signal $\mathrm{a}_{i,t}$ at any time $t$ and the corresponding generated trainable attention weight $\mathrm{w}_i$ from the previous softmax-activated layer. These trainable attention weights help the particular features to get attention that correspond to an elicited emotion. The attention weights get updated subsequently with back-propagation in the neural network and assign the appropriate weightage to the particular physiological signal, thereby giving required attention to the vital features. The second level GRUs further analyze the weighted signals and the final hidden state outputs are again concatenated in a layer. This concatenation layer is the input to an $m$-neuron dense layer which is softmax-activated to yield an $m$-class classification. The $m$-class classification corresponds to the $m$ discrete target labels for valence, arousal, and dominance ranging from 1 to $m$.

#### 3.1.1 Significance of the Biosignals in Emotion Recognition

In this work, we use an AT2GRU model which focuses on the biosignals that are more significant for a particular emotion. The significance can be understood by analyzing the different frequency bands of EEG signal and the other biosignals. The $\delta$ waves (1-3Hz) are dominant during deep sleep and these waves are slowest among all bands while having the highest amplitude. The $\theta$ waves (4-7Hz) are dominant in a relaxed state of mind that is relieved from any stress. When the mind is calm but alert, for example, during learning something, the $\alpha$ waves (8-13Hz) are significant which are slow but large in amplitude. The small and fast $\beta$ waves (14-30Hz) are significant during anxiousness or tensed state of mind. The $\gamma$ waves (31-50Hz) are the fastest and most delicate which symbolize consciousness and perception. In [14], it is shown that the $\alpha$ waves are high for high emotion as compared to a sad emotion. [15] also shows that higher frequency bands of EEG signals contain more information regarding positive emotions (high valence) than the negative emotions (low valence). A positive correlation is shown to exist between the $\beta$ waves and positive emotion [1]. [2] also reveals a strong correlation between EEG frequency bands and the corresponding valence. Also, a negative correlation is shown to exist between arousal, and the $\theta$ and $\alpha$ bands. The features from an ECG signal like heart-rate in the time domain and heart-rate variability in the frequency domain also help in recognizing the human emotion [1], [11], [30]. The heart beats faster when a person is anxious or frightened. Also, the heart-rate increases during excitement. The heart-rate variability decreases during sadness, fright, and happiness [1]. GSR gives a measure of the skin-resistance that decreases with an increase in perspiration, especially during emotions like stress or surprise [2]. Respiration and Skin Temperature (ST) also vary with different emotions. Respiration is slower during relaxation and irregular at times of anger or fear. The EMG indicates the intensity of muscle activity while the EOG gets affected by the rate of eye blinking which is correlated with anxiety [2].

Therefore, the multi-modal signals are significant differently in determination of the human emotion. As we are using a deep learning model, the predicted labels can be made to match the target labels while minimizing the prediction error by back-propagating and training for multiple epochs. Hence, the attention model learns the significant signals for a particular emotion thereby yielding a higher classification accuracy.

#### 3.1.2 Multi-class Classification using AT2GRU

The human emotion is recognized based on the value of valence, arousal, and dominance at the output. All the values of valence/arousal/dominance are discretely scaled for predicting the human emotion, such as, from 1 to 5 or 1 to 9 using the DREAMER or the DEAP dataset respectively. However, for determining the classification accuracy, we separately study the classification accuracy of valence, arousal, and dominance by comparing the predicted labels with the labels assigned by the subjects during self-

assessment. The state-of-the-art works are limited to binary classification of valence, arousal, and dominance, that is, whether is the value is low or high. A threshold label, which is mostly 3 for DREAMER and 5 for DEAP, is decided above which a value is considered as high and below as low. Also, most of these works use machine learning models with binary classifiers. However, unlike the existing works, our deep learning based AT2GRU model performs a $m$-class classification from 1 to $m$, similar to the $m$ different target labels as obtained from self-assessment of subjects. The softmax activation at the dense layer after the second level of GRUs is used for getting the multi-class classification. As the labels are categorical, a categorical cross-entropy loss function is present which gets mitigated to minimum by back-propagating during subsequent epochs.

#### 3.1.3 Computational Complexity of AT2GRU Model

For the proposed AT2GRU model in Figure 1, as we employ $g$ number of GRUs at each of the two levels for $g$ biosignals, therefore each GRU has a complexity of $O(1)$ per time step and weight [12]. Considering the input to the second level GRUs, there are $g$ multiplications between attention weights and output of first level GRUs. There are one division and $(g-1)$ additions for each of the $g$ attention weights at the softmax output after first level. Similarly, after the second level GRUs, there are again one division and $(m-1)$ additions at the softmax layer. Therefore, the overall computational complexity of the proposed AT2GRU model for $t$ time-steps per weight can be written as $O(2gt + gt + g(1 + (g-1)) + m(1 + (m-1))) \approx O(gt)$ after counting the FLOPs, with $t >> g$ and $t >> m$. Also, it can inferred that a similar single level LSTM/ GRU model with $g$ units and $m$ output classes will have a complexity of $O(gt)$ for $t$ time steps per weight.

### 3.2 Analysis of AT2GRU Model

The AT2GRU model emphasizes on the signals that are important corresponding to a particular emotion. In Figure 1, we observe that the GRUs are present at two levels of the proposed model. From each GRU of the first level, the final hidden state $h_t$ is concatenated in a layer which is softmax-activated to yield the trainable weights. The trainable weights $\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_g$ correspond to the weightage given to each of the $g$ biosignals respectively. As each weight is obtained from the last hidden state, therefore after $t$th time-step we have:

$$\begin{bmatrix}\mathrm{w}_1 \ldots \mathrm{w}_g\end{bmatrix}^\top = \frac{1}{\sum_{j=1}^{g} e^{h_{j,t}}} \begin{bmatrix}e^{h_{1,t}} \ldots e^{h_{g,t}}\end{bmatrix}^\top \tag{6}$$

where $e^{h_{j,t}}$ corresponds to signal $\mathbf{a}_j$. Therefore, the matrix of attention weights has a size of $g \times 1$ for $g$ signals. As the input to the second level GRUs is the product of each signal with the corresponding attention weight, we can write the input $\mathrm{x}_{i,t}$ for the second level GRUs at time $t$ as a Hadamard product:

$$\begin{bmatrix}\mathrm{x}_{1,t} & \ldots & \mathrm{x}_{g,t}\end{bmatrix}^\top = \begin{bmatrix}\mathrm{w}_1 & \ldots & \mathrm{w}_g\end{bmatrix}^\top \circ \begin{bmatrix}\mathrm{a}_{1,t} & \ldots & \mathrm{a}_{g,t}\end{bmatrix}^\top \tag{7}$$

The dominant signal for a particular emotion achieves the maximum weight $\mathrm{w}_i$ gradually with back-propagation. Every hidden state $h_{j,t}$ of GRU corresponding to $j$th signal is impacted by the sigmoid activated update gate $z_{j,t}$ and reset gate $r_{j,t}$. The hidden state is updated by passing on some information from the past using $z_{j,t}$ and washing out the unnecessary previous data using $r_{j,t}$. The update happens as $\hat{h}_{j,t} = \tanh(r_{j,t} \circ U h_{j,t-1} + V \mathrm{a}_{j,t})$ where $U$, $V$ are the weights, $\mathrm{a}_{j,t}$ is any of the EEG or other physiological signal inputs, and $\circ$ denotes element-wise multiplication. The Hadamard product of the sigmoid activated reset gate $r_{j,t}$ with the weighted previous hidden state $U h_{j,t-1}$ decides on which information to refrain from passing to next state. The final hidden state of a GRU is obtained as $h_{j,t} = (1 - z_{j,t}) \circ h_{j,t-1} + z_{j,t} \circ \hat{h}_{j,t}$. As the two gates are sigmoid activated, therefore, $z_{j,t}$ or $r_{j,t}$ can have a minimum value of zero and a maximum value 1. Considering that the update gate is assigned to maximum, that is, $z_{j,t} = 1$, we have the expression of hidden state reduced to $h_{j,t} = \hat{h}_{j,t}$. Therefore, when the hidden state is updated completely, we have the value of $h_{j,t}$ in the range $[-1, 1]$ as the $\tanh$ activation also has the same range. Each trainable weight at the concatenation layer output will also have a value from 0 to 1 because of softmax activation. These trainable weights are multiplied with the corresponding biosignals and the products are fed as input to the second level GRUs.

For the second level GRUs, the update and reset gates leverage the weighted input signals. The final hidden states of the second level GRUs are again concatenated and mapped into a dense layer with $m$ units. The softmax-activation applied at this dense layer to produce the predicted output labels, generates the categorical cross-entropy loss. The output is one-hot encoded with 1 assigned to the highest weighted output label and 0 assigned to rest. On back-propagating, the weights in the network get updated while minimizng the loss function. As we use GRU which is a RNN, therefore back-propagation through time happens from time $t$ to 0 in the hidden states. Similarly, all the other trainable attention weights from $\mathrm{w}_2$ to $\mathrm{w}_g$ get updated through back-propagation thereby prioritizing the biosignals for a particular emotion. This happens based on the back-propagation happening through time across the respective samples. The back-propagation goes down through the first level GRUs to the input signals. The prioritizing of input samples is the attention that enhances the classification accuracy of the proposed AT2GRU model.

#### 3.2.1 Impact of Attention Weights on Loss Minimization

The gradient of loss function with respect to the hidden state output of GRU varies with the attention weights if attention weights are used at the input of GRU. Considering a second level GRU for signal $\mathrm{a}_{i,t}$ with update gate $z_{i,t} = 1$ and reset gate $r_{i,t} = 0$, the corresponding hidden state $h_{i,t} = \tanh(\mathrm{w}_i V \mathrm{a}_{i,t})$. Although $z_{i,t}$ and $r_{i,t}$ can have any intermediate value between 0 and 1, we consider the extreme values for the ease of our calculation. However, for any other value, the same analysis is applicable. Assuming similar condition for the first level GRU corresponding to the same signal $\mathrm{a}_{i,t}$, we have $h_{i,t} = \tanh(V \mathrm{a}_{i,t})$, where $\mathrm{a}_{i,t}$ is a non-zero positive value as discussed in data preprocessing. Now, by the property of softmax function, the sum of attention weights $\sum_{i=1}^{g} \mathrm{w}_i = 1$. Therefore, we have $\mathrm{w}_i \leq 1$ from which we can write $\mathrm{w}_i V \mathrm{a}_{i,t} \leq V \mathrm{a}_{i,t}$. As $\tanh$ is a strictly monotonic increasing function, we can say:

$\tanh(\mathrm{w}_i V\mathbf{a}_{i,t}) \leq \tanh(V\mathbf{a}_{i,t})$ from where we can further write:

$$1-\tanh^2(\mathrm{w}_i V\mathbf{a}_{i,t}) \geq 1-\tanh^2(V\mathbf{a}_{i,t}) \tag{8}$$

Each side of Equation 8 is the differentiation of the $\tanh$ activation function with respect to the corresponding variable. Therefore, we have:

$$\frac{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}{\partial(\mathrm{w}_i V\mathbf{a}_{i,t})} \geq \frac{\partial \tanh(V\mathbf{a}_{i,t})}{\partial(V\mathbf{a}_{i,t})} \tag{9}$$

where $\tanh(\mathrm{w}_i V\mathbf{a}_{i,t})$ and $\tanh(V\mathbf{a}_{i,t})$ are the hidden state outputs for attention-weighted input and input without attention weight respectively. Now, considering the minimization of loss function $L_{\text{att}}$ with respect to weight $V$ of GRU when the hidden state $h_{i,t} = \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})$, we have:

$$\begin{aligned}\frac{\partial L_{\text{att}}}{\partial V} &= \frac{\partial L_{\text{att}}}{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}\frac{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}{\partial(\mathrm{w}_i V\mathbf{a}_{i,t})}\frac{\partial(\mathrm{w}_i V\mathbf{a}_{i,t})}{\partial V} \\ &= \mathrm{w}_i\mathbf{a}_{i,t}\frac{\partial L_{\text{att}}}{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}\frac{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}{\partial(\mathrm{w}_i V\mathbf{a}_{i,t})}\end{aligned} \tag{10}$$

and that when the hidden state $h_{i,t} = \tanh(V\mathbf{a}_{i,t})$, we have:

$$\begin{aligned}\frac{\partial L_{\text{no-att}}}{\partial V} &= \frac{\partial L_{\text{no-att}}}{\partial \tanh(V\mathbf{a}_{i,t})}\frac{\partial \tanh(V\mathbf{a}_{i,t})}{\partial(V\mathbf{a}_{i,t})}\frac{\partial(V\mathbf{a}_{i,t})}{\partial V} \\ &= \mathbf{a}_{i,t}\frac{\partial L_{\text{no-att}}}{\partial \tanh(V\mathbf{a}_{i,t})}\frac{\partial \tanh(V\mathbf{a}_{i,t})}{\partial(V\mathbf{a}_{i,t})}\end{aligned} \tag{11}$$

If the minimization of loss function is same in Equations 10 and 11 with respect to weight $V$, then:

$$\frac{\partial L_{\text{att}}}{\partial V} = \frac{\partial L_{\text{no-att}}}{\partial V} \tag{12}$$

Referring to Equation 9, we can rewrite Equation 10 as:

$$\frac{\partial L_{\text{att}}}{\partial V} \geq \mathrm{w}_i\mathbf{a}_{i,t}\frac{\partial L_{\text{att}}}{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}\frac{\partial \tanh(V\mathbf{a}_{i,t})}{\partial(V\mathbf{a}_{i,t})} \tag{13}$$

Considering Equations 11 and 12, we rewrite Equation 13:

$$\frac{\partial L_{\text{no-att}}}{\partial \tanh(V\mathbf{a}_{i,t})}\frac{\partial \tanh(V\mathbf{a}_{i,t})}{\partial(V\mathbf{a}_{i,t})} \geq \frac{\mathrm{w}_i\partial L_{\text{att}}}{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})}\frac{\partial \tanh(V\mathbf{a}_{i,t})}{\partial(V\mathbf{a}_{i,t})} \tag{14}$$

which can be simplified as:

$$\left(\frac{\partial L_{\text{att}}}{\partial \tanh(\mathrm{w}_i V\mathbf{a}_{i,t})} \Big/ \frac{\partial L_{\text{no-att}}}{\partial \tanh(V\mathbf{a}_{i,t})}\right) \leq \frac{1}{\mathrm{w}_i} \tag{15}$$

The range of $\frac{1}{\mathrm{w}_i}$ is $[1, \infty)$. It can be inferred from Equation 15 that the loss with respect to hidden state of GRU having attention-weighted input can vary at most $\frac{1}{\mathrm{w}_i}$ times the loss with respect to hidden state of GRU without attention-weighted input and this can be observed from Figure 3a.

#### 3.2.2 Attention Weight Generation and One-hot Encoded Multiclass Classification: An Example

Let us consider an example where $\mathbf{a}_1$ signal is dominant over five other physiological signals $\mathbf{a}_2$, $\mathbf{a}_3$, $\mathbf{a}_4$, $\mathbf{a}_5$, and $\mathbf{a}_6$. Considering full update, the hidden state representation at time $t$ for a first level GRU is $h_{i,t} = \tanh(r_{i,t} \circ Uh_{i,t-1} + V\mathbf{a}_{i,t})$. As the hidden states are $\tanh$ activated in the range -1 to 1, therefore we again assume that after an update the $\mathbf{a}_1$ signal has maximum hidden state value of 1, and other hidden states less than that. Let the exemplar representation be:

$$\begin{aligned}&\begin{bmatrix}h_{1,t} & h_{2,t} & h_{3,t} & h_{4,t} & h_{5,t} & h_{6,t}\end{bmatrix}^\top \\ &= \begin{bmatrix}1 & 0.6 & 0.3 & 0.1 & 0 & -0.4\end{bmatrix}^\top\end{aligned} \tag{16}$$

On application of softmax activation, we have the corresponding weight matrix as:

$$\begin{aligned}&\begin{bmatrix}\mathrm{w}_1 & \dots & \mathrm{w}_6\end{bmatrix}^\top = \frac{1}{\sum_{j=1}^{6} e^{h_{j,t}}}\begin{bmatrix}e^{h_{1,t}} & \dots & e^{h_{6,t}}\end{bmatrix}^\top \\ &= \frac{1}{8.665}\begin{bmatrix}2.718 & 1.822 & 1.350 & 1.105 & 1 & 0.670\end{bmatrix}^\top \\ &= \begin{bmatrix}0.314 & 0.210 & 0.156 & 0.128 & 0.115 & 0.077\end{bmatrix}^\top\end{aligned} \tag{17}$$

From the weight vector, we can find that their sum is 1 with the highest weight being $0.314$ assigned to $h_{1,t}$ corresponding to $\mathbf{a}_1$ signal. In every time-step, the attention weights are multiplied with the corresponding biosignal, thereby assigning weightage to them. The model learns by updating the weights including these attention weights during back-propagation through time.

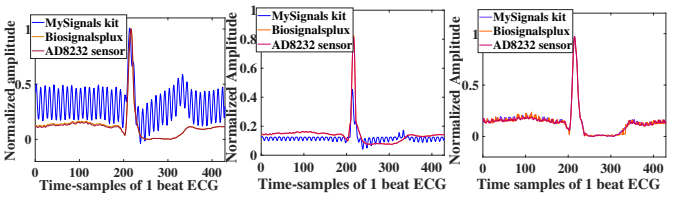
The softmax-activated output from the dense layer after the second level GRUs is one-hot encoded with the maximum weighted output assigned as 1 while the rest becomes 0. An example is shown as follows:

$$\underset{\textbf{Softmax output}}{\begin{bmatrix}0.56 & 0.31 & 0.03 & 0.04 & 0.06\end{bmatrix}} \rightarrow \underset{\textbf{One-hot encoded output}}{\begin{bmatrix}1 & 0 & 0 & 0 & 0\end{bmatrix}} \rightarrow \underset{\textbf{Label}}{\begin{bmatrix}1\end{bmatrix}}$$

## 4 Experiments and Results

We use the DREAMER and the DEAP datasets for recognition of human emotion and also show the superiority of the proposed AT2GRU model over GRU and LSTM. Unlike most of the state-of-the-art works availing the EEG data from the DREAMER dataset, we leverage both the available ECG and EEG signal recordings of the same dataset for human emotion recognition. 23 subjects including 14 males and 9 females participate in the experiment. The EEG data is acquired using Emotiv EPOC wireless EEG headset through 14 EEG electrodes at a sampling rate of 128Hz while the ECG signal is taken using the wireless SHIMMER™ ECG sensor at 256Hz [1]. 18 movie clips are shown to each of these 23 subjects for eliciting 9 different emotions namely amusement, anger, calmness, fear, excitement, happiness, disgust, sadness, and surprise. The duration of each film clip varies between 65 to 393 seconds, which is assumed to be sufficient for eliciting the emotions [1]. In order to avoid mixing of multiple emotions, the clipping of the last 60 seconds for each film is used in our experiment. Each subject is made to watch a neutral movie clip so as to get the neutral emotion state contributiong to the base values for each kind of input sample. The subjects are self-assessed at the end of each movie to scale valence, arousal, and dominance in the range of 1 to 5. In [1], where the DREAMER dataset is used first, the valence, arousal, and dominance are assigned with the binary states, that is, low and high, for predicting emotions. However, we use multi-class prediction labels from 1 to 5 corresponding to that of the self-assessment of the subjects.

In the DEAP dataset, 32 healthy participants comprising 16 males and 16 females participate in the experiment [2].




(a) (b) (c)

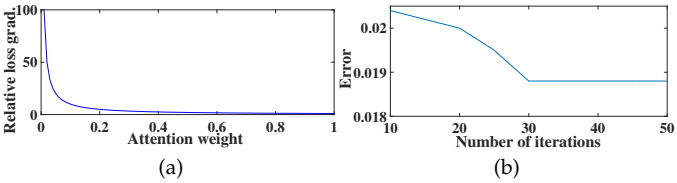
Fig. 2: 1 beat ECG signal of same person from 3 devices (a) heterogeneous (RMSE = 0.239) (b) NM method-based heterogeneity mitigated (RMSE = 0.124) (c) MDHM method-based heterogeneity mitigated (RMSE = 0.021)

Each participant watches 40 music videos each of one-minute duration. Although both the raw and pre-processed data are available in the dataset, we leverage the raw dataset and pre-process that using the same three filters like that of DREAMER. Along with EEG, the peripheral nervous system signals namely GSR, respiration amplitude, ST, blood volume, EMG, and EOG are also measured. The EEG and the peripheral signals are recorded using 32 channels and 12 channels respectively at a sampling rate of 512Hz. A trial segment of the experimental data is of 60 second duration while a baseline signal is of 3 second duration. For each video, a participant rates the valence and arousal on a continuous scale between 1 and 9. In this case also, the valence and arousal are rated as low or high based on a threshold of value 5 in the parent dataset [2]. In this work, similar to the DREAMER dataset, we perform a multi-class prediction on the DEAP dataset for valence and arousal values with the discrete labels ranging from 1 to 9.

From the EEG signal, the $\delta$, $\theta$, $\alpha$, $\beta$, and $\gamma$ wave bands are extracted for each dataset. Each of these frequency bands is used as a signal input to the AT2GRU model. Every frequency band has different significance for different emotions. Based on the significance, the attention weights prioritize the wave bands. Similarly, the other physiological signals present in each dataset are assigned with the trainable attention weights.

### 4.1 Heterogeneity Mitigation with MDHM Method

We mitigate the heterogeneity among the various biomedical devices so as to increase the consistency in the classification accuracy of the model. We train and test the LSTM, GRU, and the proposed AT2GRU models with a particular physiological signal obtained from one device. However, for the same person, if the data is acquired using some other device from a different manufacturer, the classification result can differ owing to device heterogeneity. On mitigating the heterogeneity, the results achieved for each person are more consistent. In Figure 2, we show one beat of normalized ECG signal acquired from a healthy male participant leveraging three different devices namely MySignals kit, Biosignalsplux, and AD8232 ECG sensor in the Sensor Networks Research Lab, IIT Patna. The participant's consent is taken in the informed consent form and the data acquisition is approved by the Ethical Committee of IIT Patna under the research proposal "Healthcare Data Analysis in IoT Networks". The ECG signal from each device is equi-sampled and compared for observing the heterogeneity. The device heterogeneity can be observed from Figure 2a. This heterogeneity among the three devices is mitigated by both the traditional NM method [28] and our proposed MDHM algorithm as observed from Figures 2b and 2c respectively. It can be observed that the error is mitigated more leveraging the MDHM method in comparison to the NM method. The NM method used in Figure 2b and the MDHM method in Figure 2c generate Root Mean Squared Errors (RMSE) of 0.124 and 0.021 respectively after 10 iterations. In order to optimize the computational complexity of the heterogeneity mitigation process, we iterate the proposed MDHM algorithm for an optimal number of iterations $l^{\star}$.




(a) (b)

Fig. 3: (a) Variation of relative gradient of loss with attention weight (b) Heterogeneity mitigation between 2 devices

#### 4.1.1 Selection of Optimum Number of Iterations

We iterate the proposed MDHM method for the required number of times to mitigate the device heterogeneity. The selection of the number of iterations is discussed in Algorithm 1. We mitigate the Manhattan distance among the samples at an instant leveraging the MDHM method. The proposed method attains the stopping criterion when the $\ell 1$ norm among devices stabilizes at the minimum value. This can be observed from Figure 3b where the error between 2 devices gets mitigated to the minimum value. It is observed that the error is minimized to $0.0188$ at about 30th iteration. Therefore, for this particular example the stopping criterion is attained at the 30th iteration. Referring to the proposed MDHM technique in Algorithm 1, we can infer that as soon as the difference between the sample values from any two devices at a particular instant becomes equal to that of previous iteration indicating stabilization, the iterations for heterogeneity mitigation stop thereby restricting the number of computations. Therefore, the stabilization in error is checked each time for every 2 samples.

### 4.2 Human Emotion Recognition using AT2GRU

The various wave bands of EEG signal are differently significant for each human emotion. Additionally, the heart-rate and its variability available from an ECG signal, the skin-resistance variation from GSR, respiration, ST, intensity of muscle activity available from EMG, and the rate of eye blinking from EOG are significant features for emotion recognition. The sequential deep learning models can learn the embedded features through the neural layers to recognize human emotion from the biosignals directly without manually extracting the features, unlike the machine learning models. The proposed AT2GRU is used in this work for recognizing the human emotion of multiple persons after watching different movie clippings separately.

TABLE 2: Average classification accuracies of valence, arousal, and dominance for 5 subjects on DREAMER dataset using denoised ECG and EEG data

| | | Valence (Accuracy%) | | | Arousal (Accuracy%) | | | Dominance (Accuracy%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Subject** | **Filter** | **LSTM** | **GRU** | **AT2GRU** | **LSTM** | **GRU** | **AT2GRU** | **LSTM** | **GRU** | **AT2GRU** |
| Person 2 | Daubechies (db8) | 80.00 | 86.11 | **86.67** | 73.33 | 83.33 | **93.33** | 80.00 | 77.75 | **80.00** |
| | Symlet (sym8) | 80.00 | 80.00 | **80.00** | 80.00 | 83.33 | **86.67** | 66.67 | 77.75 | **80.00** |
| | Coiflet (coif5) | 66.67 | 83.33 | **80.00** | 66.67 | 77.78 | **80.00** | 80.00 | 77.75 | **80.00** |
| Person 5 | Daubechies (db8) | 73.33 | 80.00 | **80.00** | 80.00 | 72.22 | **78.61** | 66.67 | 80.55 | **81.67** |
| | Symlet (sym8) | 66.67 | 86.67 | **83.33** | 73.33 | 69.45 | **73.34** | 86.67 | 86.11 | **86.67** |
| | Coiflet (coif5) | 86.67 | 83.33 | **90.00** | 73.33 | 72.22 | **76.67** | 73.33 | 86.11 | **83.33** |
| Person 12 | Daubechies (db8) | 66.67 | 86.11 | **86.67** | 66.67 | 83.33 | **86.67** | 80.00 | 77.75 | **78.33** |
| | Symlet (sym8) | 80.00 | 83.33 | **86.67** | 60.00 | 77.78 | **83.33** | 80.00 | 86.11 | **86.67** |
| | Coiflet (coif5) | 73.33 | 86.11 | **80.00** | 73.33 | 77.77 | **80.00** | 80.00 | 80.55 | **83.33** |
| Person 19 | Daubechies (db8) | 80.00 | 77.78 | **78.34** | 80.00 | 86.11 | **90.00** | 80.00 | 86.11 | **86.67** |
| | Symlet (sym8) | 73.33 | 80.55 | **86.67** | 73.33 | 80.55 | **86.67** | 73.33 | 69.44 | **73.33** |
| | Coiflet (coif5) | 73.33 | 83.33 | **83.33** | 73.33 | 83.33 | **73.33** | 80.00 | 83.33 | **83.33** |
| Person 23 | Daubechies (db8) | 73.33 | 83.33 | **83.33** | 86.67 | 77.78 | **85.00** | 73.33 | 88.88 | **86.67** |
| | Symlet (sym8) | 73.33 | 66.67 | **76.67** | 73.33 | 77.78 | **83.33** | 73.33 | 83.33 | **83.33** |
| | Coiflet (coif5) | 80.00 | 83.33 | **75.00** | 66.67 | 77.77 | **80.00** | 73.33 | 80.55 | **80.00** |

TABLE 3: Average classification accuracies of valence and arousal for 5 subjects on DEAP dataset using denoised physiological data

| | | Valence (Accuracy%) | | | Arousal (Accuracy%) | | |
|---|---|---|---|---|---|---|---|
| **Subject** | **Filter** | **LSTM** | **GRU** | **AT2GRU** | **LSTM** | **GRU** | **AT2GRU** |
| Person 3 | Daubechies (db8) | 73.45 | 75.33 | **79.88** | 75.67 | 77.67 | **80.33** |
| | Symlet (sym8) | 65.34 | 71.67 | **74.45** | 76.33 | 77.00 | **81.00** |
| | Coiflet (coif5) | 72.33 | 73.33 | **75.67** | 77.33 | 79.67 | **81.67** |
| Person 5 | Daubechies (db8) | 76.33 | 78.45 | **82.33** | 78.33 | 78.33 | **80.33** |
| | Symlet (sym8) | 73.33 | 77.67 | **78.67** | 77.83 | 80.67 | **82.33** |
| | Coiflet (coif5) | 71.67 | 76.67 | **77.67** | 80.67 | 80.67 | **82.67** |
| Person 8 | Daubechies (db8) | 76.67 | 77.33 | **80.33** | 81.00 | 81.33 | **84.67** |
| | Symlet (sym8) | 73.33 | 78.33 | **82.67** | 83.33 | 83.67 | **84.33** |
| | Coiflet (coif5) | 72.67 | 74.67 | **76.33** | 78.67 | 80.33 | **81.33** |
| Person 11 | Daubechies (db8) | 79.67 | 82.33 | **83.33** | 81.33 | 82.33 | **84.67** |
| | Symlet (sym8) | 82.33 | 84.00 | **84.33** | 83.67 | 84.33 | **84.33** |
| | Coiflet (coif5) | 80.67 | 81.23 | **81.67** | 79.33 | 80.00 | **81.67** |
| Person 17 | Daubechies (db8) | 76.67 | 77.33 | **79.00** | 82.33 | 84.00 | **84.33** |
| | Symlet (sym8) | 73.33 | 76.33 | **82.33** | 78.67 | 80.33 | **83.67** |
| | Coiflet (coif5) | 78.33 | 78.67 | **79.23** | 77.33 | 78.45 | **80.33** |

#### 4.2.1 Tuning of Hyperparameters

We tune the hyperparameters of the AT2GRU model to maximize the emotion recgnition accuracy. The hyperparameters that we tune are learning rate, number of hidden units, and batchsize. The learning rate for our model is tuned to 0.001. For the DREAMER dataset, the batchsize is kept as one where each movie denotes one sample of the dataset. As we consider arbitrary 10 movies watched by each of the 23 persons to train our model and the rest 8 movies for testing, therefore the maximum training data available to us is of size 23×10 movies. Similarly, for the DEAP dataset also, the batchsize is kept as unity with the training data being of size 32×30 considering 30 music videos for each of the 32 participants. For testing purpose, rest 10 movies are used corresponding to each of these 32 persons. The number of hidden units for each GRU layer is kept as 5 for the DREAMER data. Increasing the hidden units above 5 leads to overfitting of the model. For DEAP data, we tune the number of hidden units to 10.

#### 4.2.2 Performance Comparison with other RNNs

We compare the performance of the proposed AT2GRU with two other sequential models namely LSTM and GRU. Leveraging the DREAMER dataset, the models are trained with the biosignals corresponding to 10 movie clippings for each of the 23 persons. The biosignals for the rest 8 movies are used in testing the models. For the DEAP dataset, the biosignals generated from 30 music videos for each of the 32 participants are used in training the models and that generated from the rest 10 music videos are used in testing purpose. The test accuracies obtained for few persons from the DREAMER and the DEAP datasets are shown in Tables 2 and 3 respectively. We obtain the denoised signals from three different filters namely db8, sym8, and coif5 as they have similar properties [29]. However, the classification performance from each of them differs owing to the variation of mean square error and signal-to-noise ratio [29]. Therefore, we consider the filter output yielding the highest accuracy for each person. It is observed that the proposed AT2GRU model outperforms the LSTM and GRU in terms of classification accuracies. The emotion recognition results obtained by us is a 5-class classification and a 9-class classification on valence/arousal/dominance for the DREAMER and the DEAP datasets respectively unlike the state-of-the-art works that perform binary classification (high/low) based on a threshold. During training, we obtain training and validation accuracies ranging from 96 percent to nearly 100 percent for every person of each dataset after running 150 epochs. We perform the test on the trained models and obtain the average performance metrics of accuracy, precision, recall, and F1-score which are tabulated in Tables 4 and 5 corresponding to the DREAMER and the DEAP datasets respectively. The average performance metrics are obtained by averaging the results of each person in the test-set. The mean accuracy values along with the Standard Deviations (SD) are shown in Tables 4 and 5. We also perform the $t$-test considering a significance level of $0.05$ and test the statistical significance of the results. The F1-score can be obtained from the harmonic mean of precision and recall values. From precision, the fraction of correctly classified labels out of the total predicted labels for a particular class can be analyzed whereas recall gives the number of correctly classified labels out of total actual labels of a class. Using the mean accuracy values and SDs of valence/arousal/dominance for any two models (LSTM-AT2GRU or GRU-AT2GRU) from Table 4 or 5 of the revised manuscript, we calculate the $t$-value $t_{\text{calc}}$. Then we compare the $t_{\text{calc}}$ with the critical $t$-value $t_{\text{crit}}$ in the $t$-distribution table corresponding to the significance level 0.05 and observe the $p$-value. The two sample $t$-test results for valence/arousal/dominance of both datasets are tabulated in Table 6. We observe from Table 6 that for all cases, $t_{\text{calc}} > t_{\text{crit}}$ resulting in $p < 0.05$. Therefore, the results observed from Tables 4 and 5 are statistically significant. The multi-class classification results obtained by us on the basis of valence/arousal/dominance help to recognize the exact emotion of a person.

#### 4.2.3 Comparison with the State-of-the-art Works

We compare the performance of the proposed AT2GRU model with the different state-of-the-art works in literature. Table 7 highlights the accuracies and the complexities of the state-of-the-art works that leverage the same datasets as ours. To the best of our knowledge, the existing works underwent bi-class and tri-class classification in terms of valence/arousal/dominance. As observed from Table 7, both machine learning and deep learning algorithms are used for human emotion recognition [1], [2], [31], [32]. The use of convolutional and recurrent neural networks with different architecture is preferred by the researchers

TABLE 4: Performance comparison of different sequential models on DREAMER dataset with multi-class classification

| | Valence | | | | Arousal | | | | Dominance | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Accuracy±SD %** | **Precision** | **Recall** | **F1-score** | **Accuracy±SD %** | **Precision** | **Recall** | **F1-score** | **Accuracy±SD %** | **Precision** | **Recall** | **F1-score** |
| **LSTM** | 76.52±3.24 | 1.00 | 0.33 | 0.50 | 75.67±4.05 | 0.50 | 0.50 | 0.50 | 77.74±3.70 | 0.50 | 0.50 | 0.50 |
| **GRU** | 82.52±0.49 | 0.66 | 1.00 | 0.80 | 83.22±0.52 | 0.50 | 0.50 | 0.50 | 83.55±0.46 | 1.00 | 0.50 | 0.67 |
| **AT2GRU** | **84.15±0.54** | **1.00** | **0.66** | **0.80** | **84.69±0.47** | **0.66** | **1.00** | **0.80** | **85.81±0.96** | **1.00** | **0.66** | **0.80** |

TABLE 5: Performance comparison of different sequential models on DEAP dataset with multi-class classification

| | Valence | | | | Arousal | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Accuracy±SD %** | **Precision** | **Recall** | **F1-score** | **Accuracy±SD %** | **Precision** | **Recall** | **F1-score** |
| **LSTM** | 75.83±2.92 | 0.75 | 0.71 | 0.73 | 79.67±1.74 | 0.77 | 0.76 | 0.76 |
| **GRU** | 78.67±1.22 | 0.77 | 0.73 | 0.75 | 81.80±0.44 | 0.83 | 0.80 | 0.81 |
| **AT2GRU** | **82.33±1.12** | **0.86** | **0.82** | **0.84** | **83.67±0.77** | **0.85** | **0.82** | **0.83** |

TABLE 6: Two sample $t-$test results for DREAMER (DRE.) and DEAP datsets corresponding to degrees of freedom (df)

| **Data** | **df** | $t_{\text{crit}}$ | **LSTM-AT2GRU ($t_{\text{calc}}$)** | | | **GRU-AT2GRU ($t_{\text{calc}}$)** | | |
|---|---|---|---|---|---|---|---|---|
| | | | **Val.** | **Arou.** | **Dom.** | **Val.** | **Arou.** | **Dom.** |
| DRE. | 44 | 2.015 | 2.32 | 2.21 | 2.11 | 2.23 | 2.10 | 2.12 |
| DEAP | 62 | 1.999 | 2.08 | 2.11 | - | 2.21 | 2.10 | - |

TABLE 7: Comparison of the state-of-the-art works on DREAMER and DEAP datasets in terms of valence (V), arousal (A), dominance (D)

| **Data-set** | **Reference** | **Classification Model** | **Classes** | **Accuracy/SD%** | **Complexity** |
|---|---|---|---|---|---|
| **DREAMER** | Katsigiannis et al. [1] | RBF-based SVM | 2 | 62.49 (V)<br>62.32 (A)<br>61.84 (D) | $O(kd)$ where $k =$ no. of support vectors, $d =$ the input dimensionality |
| | Song et al. [8] | DGCNN | 2 | 86.23/12.29(V)<br>84.54/10.18 (A)<br>85.02/10.25 (D) | $O(\Sigma_{i=1}^{d} b_{i-1} s_i^2 k_i l_i^2)$ where $d, k, b, s, l$ denote no. of layers, filters, input channels, spatial size of filter, and spatial size of output feature map |
| | **Our work** | LSTM | **5** | 76.52/3.24 (V)<br>75.67/4.05 (A)<br>77.74/3.70 (D) | $O(gt)$ for $t$ time steps per weight with $g$ signals |
| | | GRU | | 82.52/0.49 (V)<br>83.22/0.52 (A)<br>83.55/0.46 (D) | |
| | | **AT2GRU** | | **84.15/0.54** (V)<br>**84.69/0.47** (A)<br>**85.81/0.96** (D) | |
| **DEAP** | Koelstra et al. [2] | Gaussian Naive Bayes | 2 | 57.60 (V)<br>62.00 (A) | $O(nK)$ where $n, K$ denote no. of features & classes |
| | Pandey et al. [31] | DNN | 2 | 62.50 (V)<br>61.25 (A) | $O(nt \circ (ij + jk + ...))$ where $n$ is number of epochs, $t$ is training examples, and $i, j, k$ are the layers |
| | Pandey et al. [32] | CNN | 3 | 61.50 (V)<br>58.50 (A) | $O(\Sigma_{i=1}^{d} b_{i-1} s_i^2 k_i l_i^2)$ where $d, k, b, s, l$ denote no. of layers, filters, input channels, spatial size of filter, and spatial size of output feature map |
| | Siddharth et al. [6] | LSTM | 2 | 79.52/0.70 (V)<br>78.34/0.69 (A) | $O(1)$ per time step and weight for each unit |
| | Chen et al. [24] | LSTM | 2 | 63.70 (V)<br>61.90 (A) | $O(1)$ per time step and weight for each unit |
| | | H-ATT-BGRU | | 67.90 (V)<br>66.50 (A) | |
| | Joshi et al. [33] | Bi-LSTM | 2 | 73.50 (V)<br>75.00 (A) | $O(1)$ per time step and weight for each unit |
| | Ma et al. [23] | LSTM | 2 | 89.04/5.79 (V)<br>91.00/4.10 (A) | $O(1)$ per time step and weight for each unit corresponding to each modality |
| | | Residual LSTM | | 91.85/1.81 (V)<br>92.54/2.18 (A) | |
| | | MMResLSTM | | 92.31/1.55 (V)<br>92.87/2.11 (A) | |
| | **Our work** | LSTM | **9** | 75.83/2.92 (V)<br>79.67/1.74 (A) | $O(gt)$ for $t$ time steps per weight with $g$ signals |
| | | GRU | | 78.67/1.22 (V)<br>81.80/0.44 (A) | |
| | | **AT2GRU** | | **82.33/1.12** (V)<br>**83.67/0.77** (A) | |

owing to the model's ability of analyzing the deeply embedded features. However, the CNN-based models have more computational complexity as compared to the sequential counterpart. Therefore, although few works like [8] achieve high accuracy leveraging CNN, they have a high computational complexity. On the other hand, the sequential models involving LSTM, bidirectional LSTM, GRU, attention GRU and few other variants have a less time complexity of $O(1)$ per time step and weight for each unit [6], [23], [24], [33]. [23] achieves a high accuracy using LSTM and its variants but for a binary classification. An attention-based bidirectional GRU model is discussed in [24] to show improved performance over LSTM. The architecture discussed in [24] comprises two levels of bidirectional GRUs namely, a sample encoder level and an epoch encoder level to create sample attention and epoch attention respectively. The epochs and the samples in each epoch for one entire EEG signal are considered by [24] unlike ours where the different frequency bands of EEG along with other biosignals are considered. Further, in their model the attention is developed at each encoder level based on the outputs from all the hidden states of the bidirectional GRUs. On the other hand, in our AT2GRU model, the last hidden state output from each GRU corresponding to a particular biosignal is considered to develop the attention based on signal's significance for emotion prediction. Additionally, we carry out a multi-class classification using the proposed AT2GRU model unlike the others and obtain close classification accuracies with the existing works that yield high accuracies.

## 5 Conclusion

In this work, we propose the MDHM method to mitigate the device heterogeneity that arises due to the varying specifications of different biomedical devices measuring the same physiological signal. Mitigation of device heterogeneity enhances the consistency in classification accuracies of biosignals acquired from devices of different make. Further, we recognize human emotion in terms of valence, arousal, and dominance values leveraging the different physiological signals. The proposed AT2GRU model performs a multi-class classification of valence, arousal, and dominance with average testing accuracies of 84.15 percent, 84.69 percent, and 85.81 percent respectively on the DREAMER dataset. Leveraging the DEAP dataset, the same AT2GRU model yields average testing accuracies of 82.33 percent and 83.67 percent for valence and arousal respectively. The labels for multi-class classification correspond to the self-assessment grading scale ranging from 1 to 5 (for DREAMER) and 1 to 9 (for DEAP) given by the subjects based on their emotions after watching the movie clippings. We perform the multi-class classification on both the datasets using two other

RNNs namely LSTM and GRU. We show that AT2GRU model surpasses them in recognition of human emotion.

In few works performing binary classification of valence/arousal/dominance, the accuracies are higher than the AT2GRU at the expense of higher computational complexity. Although we carry out a multi-class classification of valence, arousal, and dominance, further research is required to improve the multi-class classification accuracy.

## REFERENCES


[1] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Bio. Heal. Info.*, vol. 22, no. 1, pp. 98–107, 2017.

[2] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2011.

[3] T. Zhang, X. Wang, X. Xu, and C. P. Chen, "GCB-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, 2019.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pat. Anal. M. Intel.*, vol. 31, no. 1, pp. 39–58, 2008.

[5] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, 2019.

[6] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, 2019.

[7] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch: An open source library for sentence-based emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 4, no. 3, pp. 312–325, 2013.

[8] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, 2018.

[9] T. Song *et al.*, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12 177–12 191, 2019.

[10] J. Cheng *et al.*, "Emotion recognition from multi-channel EEG via deep forest," *IEEE J. Bio. Heal. Info.*, 2020.

[11] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, 2011.

[12] P. B. Weerakody, K. W. Wong, G. Wang, and W. Ela, "A review of irregular time series data handling with gated recurrent neural networks," *Neurocomputing*, vol. 441, pp. 161–178, 2021.

[13] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[14] A. Ashtaputre-Sisode, "Emotions and brain waves," *The Int. J. Ind. Psyc.*, vol. 3, no. 2, pp. 14–18, 2016.

[15] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Aff. Comput.*, vol. 7, no. 1, pp. 17–28, 2015.

[16] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Int. Conf. M. Learn.* PMLR, 2016, pp. 2397–2406.

[17] A. Stisen *et al.*, "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," in *Proc. ACM Conf. Emb. Net. Sen. Sys.*, 2015, pp. 127–140.

[18] A. M. Abdelmoniem and M. Canini, "Towards Mitigating Device Heterogeneity in Federated Learning via Adaptive Model Quantization," in *Proc. Work. M. Learn. Sys.*, 2021, pp. 96–103.

[19] U. S. F. D. Administration, "Understanding barriers to medical device quality," *US Food and Drug Administration: Silver Spring, MD, USA*, 2011.

[20] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intel. Sys. Tech.* , vol. 3, no. 3, pp. 1–30, 2012.

[21] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Trans. Affect. Comput.*, 2020.

[22] E. H. Siegel *et al.*, "Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories," *Psycho. Bul.*, vol. 144, no. 4, p. 343, 2018.

[23] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 176–183.

[24] J. Chen, D. Jiang, and Y. Zhang, "A hierarchical bidirectional GRU model with attention for EEG-based emotion classification," *IEEE Access*, vol. 7, pp. 118 530–118 540, 2019.

[25] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel, "BioCNN: A hardware inference engine for EEG-based emotion detection," *IEEE Access*, vol. 8, pp. 140 896–140 914, 2020.

[26] M. Huang *et al.*, "A clinical decision support framework for heterogeneous data sources," *IEEE J. Bio. Heal. Info.*, vol. 22, no. 6, pp. 1824–1833, 2018.

[27] S. Kumar and S. K. Das, "ZU-mean: fingerprinting based device localization methods for IoT in the presence of additive and multiplicative noise," in *Proc. Work. Prog. Int. Conf. Dist. Comput. Net.* ACM, 2018, p. 15.

[28] S. Singer and J. Nelder, "Nelder-mead algorithm," *Scholarpedia*, vol. 4, no. 7, p. 2928, 2009.

[29] M. K. Wali, M. Murugappan, and B. Ahmmad, "Wavelet packet transform based driver distraction level classification using EEG," *Math. Prob. Eng.*, vol. 2013, 2013.

[30] M. K. Abadi *et al.*, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, 2015.

[31] P. Pandey and K. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," *J. King Saud Univ.-Comput. Info. Sc.*, 2019.

[32] P. Pandey and K. Seeja, "Subject independent emotion recognition system for people with facial deformity: an eeg based approach," *J. Amb. Intel. Hum. Comput.*, pp. 1–10, 2020.

[33] V. M. Joshi and R. B. Ghongade, "IDEA: Intellect database for emotion analysis using eeg signal," *J. King Saud Univ.-Comput. Info. Sc.*, 2020.


**Pritam Khan** received the M.E. degree from Jadavpur University, Kolkata, India in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Indian Institute of Technology Patna, Patna, India. His broad research interests lie in the field of IoT, machine learning, deep learning, cyber-physical system, and smart healthcare.

**Priyesh Ranjan** received the B.Tech degree from the Department of Electrical Engineering, Indian Institute of Technology Patna, Patna, India. His broad research interests lie in the field of IoT, machine learning, deep learning, cyber-physical system, and smart healthcare.

**Sudhir Kumar** (Senior Member, IEEE) received the Ph.D. degree from the Electrical Engineering (EE) Department, Indian Institute of Technology Kanpur, Kanpur, India, in 2015. He is currently an Assistant Professor with the EE Department, Indian Institute of Technology Patna, Patna, India. He published more than 50 research articles in prestigious journals and conference proceedings. His broad research interests include wireless sensor networks and Internet of Things.