DREAM: Deep Learning-based Recognition of Emotions from Multiple Affective Modalities using consumer-grade body sensors and video cameras

Aditi Sharma, *Member, IEEE*, Akshi Kumar, *Senior Member, IEEE*

***Abstract*— Ambient smart cities exist on the intersection of digital technology, disruptive innovation and urban environments that now essentially augment affect empathy and intelligent interfacing for human computer interactions (HCI). This research puts forward a deep learning approach, DREAM, for recognition of emotions using three affective modalities (audio, video, physiological) to develop an empathetic HCI system using consumer electronic IoT sensors and cameras. Convolution network is used to train for physiological signals. VGG and ResNet have been used to pre-train the models for emotion recognition from video and audio signals. DREAM is then fine-tuned on the publicly available K-EmoCon dataset to accurately recognize emotion for each subject. K-EmoCon is annotated by seven persons for five discrete emotions, and two affect dimensions. Finally, a probability-based average decision-level fusion strategy is used for combining the outputs of all the modalities. Leave one out strategy is used to train and evaluate the model for subject specific accuracies. For discrete emotions highest accuracy of 81.7% and 82.4% is achieved for dimensional emotions. DREAM has performed better than existing state-of-the-art for both emotion models.**

***Index Terms*— Facial Emotion Recognition, Human Computer Interaction, IoT sensors, K-EmoCon, Multi Modalities, Transfer Learning.**

I. INTRODUCTION

AMBIENT Intelligence (AmI) enriches the urban spaces as it combines end-point hardware, software, user experience, machines/human machine interaction and learning. Smart technology, data, sensors, video camera, apps, and citizen science, all have the potential to influence and shape urban environment in an ambient smart city. Built upon the advances in sensors and sensor networks, cameras, pervasive computing, and artificial intelligence, an ambient smart city essentially includes systems characterized as sensitive, responsive, adaptive, transparent, ubiquitous, and intelligent [1,2]. But a city is not a machine and is rather made by citizen's actions and feelings. Augmented intelligence [3] (also known as intelligence amplification, or IA), a subsection of artificial intelligence (AI) has emerged as a symbiotic relationship between man and machine to enhance human intelligence within these ambient smart cities. IA has been part of the human-centered partnership model of people and AI for years. As the IoT and smart objects(camera) connectivity expands, we can expect to see IA in almost all aspects of life to enhance cognitive performance, including learning, decision making and new experiences that can lead the way to develop safe and reliable work place for humans in industry 5.0[4].

Emotions are an inseparable aspect of human intelligence and integral to decision making. Researchers are working in the field of human computer interaction (HCI) for more than a decade now, and emotion recognition and understanding of human senses (like taste and smell) remain critically important for HCI [7]. Chatbots and AI-powered conversational interfaces are reshaping the world of HCI. Moreover, most ambient-assisted applications in smart cities and industries still lack the understanding of human sentiments and emotions [5].

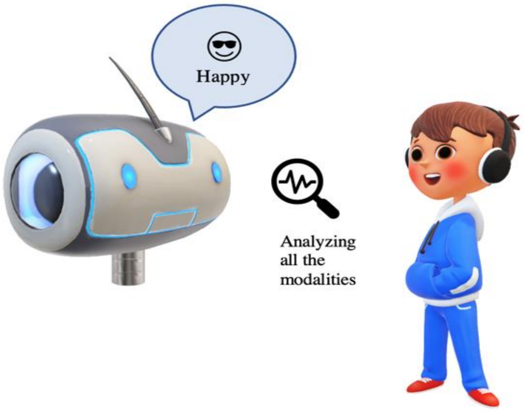

Fig.1. Human Computer Interaction

An obvious solution is to integrate the intelligent amplification of things (IAoT) with emotion sensing which can have a variety of applications such as, active, and assisted living, health care and industry. It aims to present an opportunity to developers for creating automated systems which can interrelate with the affective and cognitive state of humans to make better decisions [6]. For example, an IAoT-enabled bot can layer on adaptations using the user's state of

Corresponding author: Akshi Kumar.

Aditi Sharma was Research Scholar at the Delhi Technological University, New Delhi, India. She is now an Assistant Professor with Thapar Institute of Engineering and Technology, Patiala, India. (e-mail: aditi.sharma@thapar.edu).

Akshi Kumar is with the Department of Computing, Goldsmiths, University of London, United Kingdom. (e-mail: akshi.kumar@gold.ac.uk).

mind and have empathy to enact like humans to operate in any client-oriented business for example, smart hotel or hospital management system or call center and business process outsourcing (BPO) service provider [8], [9] using consumer grade cameras and sensors. Due to easy availability of these cameras and sensors in current era, it has the potential to revolutionize the way we understand and respond to human emotions, particularly in the field of consumer behavior. With the use of these sensors and cameras, it is possible to gather data on the emotional states of individuals in real-time, which can then be used to inform marketing strategies, product design, and customer service [10].

The role of empathy has come to prominence in HCI as the community increasingly engages with issues in medical, health and emotionally charged contexts [11]. As the natural human-human interaction (HHI) is fundamentally multimodal, this fosters the need to exploit multiple affective modalities to build an empathetic HCI (Fig.1).

Pertinent literature shows the use of a variety of modalities ranging from facial expressions, body gestures, speech, and physiological signals [12], [13], [14] for sensing affective states. Recent research also advocates that a reliable automatic affect recognition system should attempt to combine two or more modalities [15, 16].

This research puts forward a deep learning approach for recognition of emotions using three affective modalities. The approach can be integrated to both AI-powered conversational chatbots and sensors, steering to an emotion-aware ambient smart city. One of the key benefits of using consumer-grade sensors and cameras for the automatic recognition of emotions is that they are relatively inexpensive and easy to use. This means that businesses of all sizes can take advantage of these technologies to improve their understanding of consumer behavior. For example, a small business owner could use facial recognition software to track customer reactions to different products or promotions, while a larger corporation could use body sensors to gather data on how customers respond to different types of advertising. The proposed approach, DREAM is built using decision-level fusion of three different affective modalities, namely, audio (speech), video (face expressions) and physiological signals. Transfer learning is used to learn from existing data. VGG and ResNet have been used to pre-train the model, followed by fine tuning using CNN for audio signals. Video signals are processed through ResNet twice for facial emotion detection and is fine-tuned on K-EmoCon using convolution network. Physiological signals are processed with help of convolution models 1-D and 2-D. The output of all the modalities is combined using probability-based average decision fusion, it is one of the techniques of late fusion [15]. The publicly available dataset for emotion recognition in conversations, K-EmoCon is used to train and test the approach.

K-EmoCon is different from other available datasets in terms of annotation of the dataset. Most available datasets for emotion recognition are either self-annotated by the subject, or by an external observer like in case of IEMOCAP. But in K-EmoCon, each instance has been annotated by 7 people, one the person himself, second the opponent/partner in the debate, and lastly by 5 external observers. And all these seven annotations are taken both in terms of discrete emotion states, and the dimensional states, the two types of emotion models [16]. Processing the different output classes available is a complex task, most of the work carried out by researchers either use self-annotation as the output label, stating a person knows best but he or she is feeling [17]. Others have used annotation by the opponent, for interaction problems, how the other person is observing, should be observed by the robot for HCI [18]. But, to include every perspective, we have processed all seven emotion annotations in both discrete form as well as dimensional form for accurate emotion detection from all perspectives.

For understanding of human emotions, different researchers have used different types of data, like IEMOCAP, MELD, CASE, CK+ for facial expressions; Berlin, LEESD for emotion from Speech. Different Modalities has been explored by researchers for emotion recognition; some have focused on facial expressions, others on hand gestures, some have worked on emotion recognition from speech, while others have focused on physiological changes. Combination of two or more modalities has been employed by researchers on some datasets [15] [16]. Saha et al used kinetic sensors for recognizing hand gestures for emotion recognition into 5 emotion states using K-NN, SVM, NN [12]. Wei et al. used Attention based CNN for emotion identification from gestures in the video signals [13].

Bertero et al. in 2017, used TED talks for emotion identification from Speech signals using CNN [14]. Lalitha et al. used different perceptual features like MFCC, PLPC with deep neural network for speech emotion recognition from Berlin dataset [19]. Fan et al. in 2021 created pre-tuned models for emotion recognition from speech signals for HCI on LSSED dataset, the largest dataset available for speech emotion recognition [20]. Kumar et al. used experimented on IEMOCAP, MELD for emotion recognition using facial expressions [15]. Chowdhary et al. in 2021 used ConvNet on pre-trained models for identifying emotions from facial expressions in CK+ dataset [21]. Ahmed et al. and Zhang et al. have worked on emotion recognition from speech signals, by analysing the linguistic parameters, and the melogram pattern of the speech signals [32,33].

K-EmoCon dataset was published in 2020, only few researchers have worked on K-EmoCon till now, none of whom used all the modalities of the dataset for emotion recognition. Gupta et al. [22], Zitouni et al. [23], Alskafi. Et al. [24], and Dissanayake et al. [25] have used only physiological signals for identifying valance-arousal emotion state of the subject. Quan et al. used only audio and visual signals of 16 participants for emotion recognition [18]. Yang et al. has used CNN on K-EmoCon for identifying emotion recognition using physiological signals only, not the face expressions or linguistic patterns at all [33]. Alhussein used CNN on only linguistic and speech signals of K-EmoCon to identify the emotions in arousal and valance dimensional emotions, having state of the art f1-score of 82% [34].

The primary contribution of the proposed DREAM is:

- Three affective modalities (audio, video, and physiological) have been used for emotion recognition.
- Different deep neural networks trained for each modality of K-EmoCon.
- For accurate emotion identification, three annotations, namely- self, partner, & externals are used.
- The performance is evaluated for both emotion models; discrete emotion states and affect dimensions.

Next section of this paper contains description of the dataset. Section 3 contains methodology used, followed by discussion of result and analysis.

II. Dataset

K-EmoCon: A multimodal publicly available dataset for emotion recognition in conversations provided by Park et al. in 2020 [26]. The dataset contains audio, video and bio signals of 32 subjects, who participated in a debate task on a social issue in teams of 2, while wearing Physiological signal measuring devices, Emperica E4, NeuroSky, and Polar H7, along with Video cameras for recording the facial expressions and gestures of the participants. On average, 10 minutes of debate was conducted between each pair, for emotion recognition, audio and video was recorded of this debate, along with Bio-signals from 3 wearable devices. ACC (32Hz), BVP (64 Hz), EDA (4Hz), HR(1Hz), IBI, and Temp (4Hz) was measured through Emperica E4; Attention, Brainwave – EEG (125Hz) and Meditation through NeuroSky; and ECG (1 Hz) from Polar H7.

K-EmoCon is the first publicly available dataset that contains natural conversation between non-actors, while monitoring their physiological changes. Another unique feature of K-EmoCon is that each instance has been annotated by 7 people, one the person himself, other by the opponent/partner in the debate, and lastly by 5 external observers. Although it makes processing the dataset difficult, but it provides different perspective for accurate emotion recognition. Park et al. have used 20 different types of emotions for annotation, including dimensional emotional model (Arousal and Valance), Basic Emotions (Cheerful, Happy, Angry, Nervous, and Sad), along with both common and less common BROMP (Baker Rodrigo Ocumpaugh Monitoring Protocol) Affective categories [26].

A. *Data Selection*

K-EmoCon contains multiple files for each modality, not all the files are accessible or are having missing entities, data cleansing was performed. The audio recording of the dataset contains 16 files, each for the debate between the participants, so each file contains audio of 2 participants. Video Signals were recorded through separate cameras for each participant, but 11 of the participants didn't allow the data for public availability, we have only access of 21 participants visual signals. Bio-signal measuring devices too had some troubles. For our proposed model, we have taken audio, video, BVP, EDA, HR, IBI, Temp, and EEG signals for accurate Emotion recognition. As only 18 participant's complete data were available for all these signals, the empirical analysis has been conducted on 18 participants data only.

B. *Pre-processing*

For emotion recognition, classification of emotions has been taken into 7 types, 2 affective dimensions and 5 emotion states as shown in fig. 2. Emotion annotation of these 7 categories were ranked on scale 1-5 for dimensional, and 1-4 for emotions by all seven annotators.

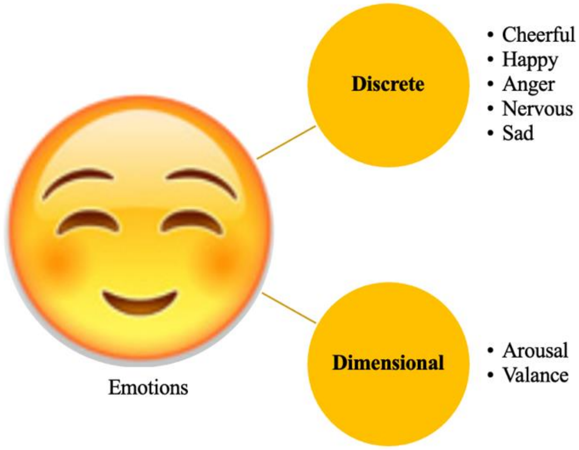

Fig. 2. Emotional Models

For our analysis we divided each emotion and dimension into two parts, low arousal if average score of all 7 annotators between 1-3, and high arousal if between 3-5, similarly very cheerful for average score of 2-4, and less cheerful for 1-2.

The experiment was evaluated for both Affective dimensions, and Emotion states separately, with 4 and 10 output classes respectively. To create time synchronous data for proposed work, the 18 participants' all signals were mapped with output (Dimensions, and Emotions), so each signal is annotated at 5 second interval with one of the four affective dimensions for first experiment, and similarly with one of the ten emotion state for second experiment.

All input files of every modality are mapped with the averaged annotated emotion. Speech signals of 22Khz sampling is cut at interval of 5 seconds and mapped to corresponding annotated emotion state. Similarly, visual signals sampled with 30 fps with frames resized into 112*112 pixels is mapped to 5 second annotated emotion by mapping 150 (30*5 frames) to single emotion instance. The bio-signals after pre-processing, is mapped to output emotion class, by using statistical feature of mean for taking 5 second data as an instance.

The bio-signals and visual signals are recorded for each participant separately, but the audio signals are recorded in sharing for the team of 2 participants. The audio signals of the opponent can trigger certain emotions in the subject, but it cannot reflect the emotions of the subject, so while pre-processing, we manually generate audio file for each participant from original file. First, a duplicate copy of the file was created for the opponent, and then manually the voice of the opponent was identified and removed from the file (sections) and replaced with the voice signals of the participant at start of the debate. Similar process was followed for opponent copy. For bio-signals, in pre-processing, missing values were replaced by means of the next two consecutive instance values.

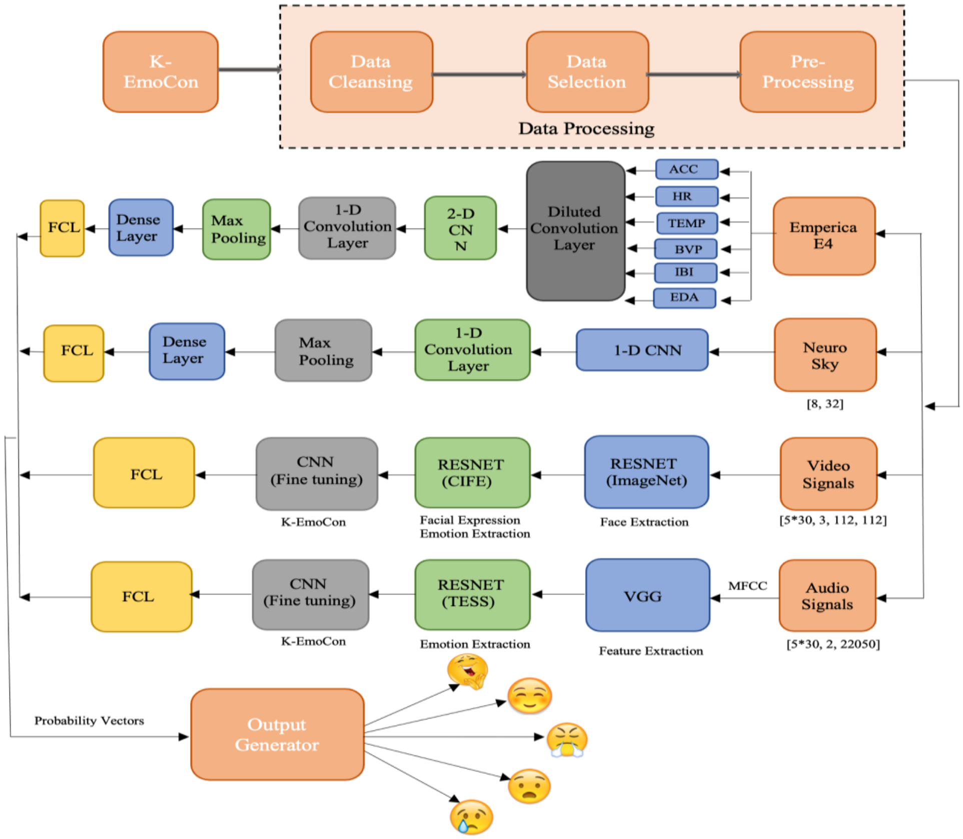

Fig.3. Architecture of DREAM

III. DREAM: Proposed Model

To create a human computer interaction system using consumer grade camera and sensors, we proposed a model with 3 modalities- DREAM. The architecture of DREAM is shown in fig 3. Initially nine input files from K-EmoCon dataset were provided to proposed model after pre-processing. Six out of nine files were of Emperica E4, Only EEG file taken from NeuroSky, Polar_H7's data had many missing values, and Heart rate has already been received from E4, so no file from Polar_H7 was processed. Video file of each participant, and Separate audio file generated after pre-processing for the participant was used as the initial input to the model.

A. *Transfer Learning*

The use of deep learning models has advanced many application areas as they provide state-of-the-art results. The models are especially useful for feature extraction owing to the ability to learn representations that cannot be modelled manually. But these models require large training datasets to perform well in real-time and building a model from scratch is time consuming and expensive in terms of data collection/labelling, privacy and training time. Transfer learning has emerged as a solution for developing and training deep neural models with less data and compute power. Transfer learning is about "transferring" the learnt representations to another problem. With transfer learning, we basically try to exploit what has been learned in one task to improve generalization in another.

For the task of emotion recognition, we follow an inductive transfer learning approach. A general representation is pre-trained on a large unlabelled dataset and is then adapted to a supervised target task using the labelled data available. The pre-trained model can also be made available for use to others who can fine-tune it for their target downstream tasks. Thus, while pre-training on a large dataset is computationally expensive, it only needs to be done once. Compared to it, downstream fine-tuning is much cheaper.

B. *Audio Modality*

After generating separate audio files for each participant, the speech signals of sampling rate 22KHz, was provided as an input ([T, C, F]), where T represents time i.e., 5s, C represents Channel i.e., 2, F represents frequency 22050Hz. Mel-

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2023.3325317

Frequency Cepstral Coefficients (MFCC) waveform encoding was used to extract audio features. For mapping audio to visual signals at same time instance, we reframed the audio signal size to that of visual signals, so temporal dimension after MFCC of audio signal became 150 (30*5). These extracted features are then provided as an input to VGG a pre-trained model [8] followed by ResNet [27] for pre-training the model for emotion recognition on TESS (Toronto Emotional Speech set) dataset. VGG creates a simple but unified deepen structure of the network, ResNet uses residual learning, that ease the training process [28]. These pre-trained models are then fine-tuned by using K-EmoCon for mapping the output to 10 emotion states. The output of fine-tuned layer is then passed to a fully connected layer (FCL).

C. *Video Modality*

After mapping emotional state at each segment of 5 second interval, the size of individual video clip became [5 x 30, 3, 112, 112] represented as ([T x F, C, W, H]), where T is time, F for frames per second, C for color channels, W and H stands for width and height respectively.

This input feature vector is feed to pre-trained ResNet on ImageNet for face extraction. The output of this ResNet model is then given to second pre-trained ResNet model on CIFE for emotion mapping of facial expressions. These pre-trained models have different with output classes, so it is then fine-tuned on the dataset itself. These visual signals can be used to identify the body gestures such as hand movements, but as the movement of participants is limited (seated on a chair), we just took facial expressions into consideration for emotion classification. Finally, the output from fine-tuned model is feed into a fully connected layer to predict the emotion state.

D. *Physiological Modality*

Seven Physiological signals have been measured by 2 wearable sensor devices Emperica E4, and NeuroSky. The band pass filtering was applied on EEG signals (8 bathes, 32 Hz), with a window size of 5 seconds to map the emotion state [29]. The input is then provided to a convolution layer with kernel of size 3 * 1, with 64 kernel and the pooling size 3*3. The output of pooling layer is again provided to 2D convolution layer, followed by max pooling.

The output of this pooling layer is provided to dense layer followed by a fully connected layer. For 6 bio signals of Emperica E4 wrist band, as the size of each signal is different, like ACC is 3*126 (for each participant), and HR 1*126, initially a diluted convolution layer is applied to all 6 bio-signals, then after generating the same size feature map, it is provided as an input to two 1D convolution layers similar to EEG signal, followed by Dense layer, and fully connected layer to produce emotion state.

E. *Convolution Neural Network*

To project the input sequences into a semantic space with a unified dimension, a 1D convolution layer or 2D convolution layer are applied on feature dimensions with varying kernel size as described in previous subsections. In the CNN layers and fully connected layers, the activation function of the Rectified Linear Unit (ReLU) is set, which handles a threshold of 0 for the negative values. This $ReLU(x)$ function is calculated as equation (1):

$$f(x) = \max(0, x_i) \tag{1}$$

The max-pooling layers are alternated between the CNN layers because of convolutional region that can increase the robustness of the features and reduce the dimensionality of the different signals vector. As a regularization technique to decrease the overfitting in the layers of the neural network, the dropout with a value of 0.5 is added. The output layers of the fully connected network are configured with the SoftMax classifier, with the purpose that the hidden layers verify the probability of predicting the emotion [35]. During the supervised training, the loss is minimized with the Root Mean Square Propagation (RMSProp) optimizer, since it adjusts the learning rate adaptively. Initially, the learning rate is set to 0.001. Once the model is executed, the knowledge base is consolidated between the vector of physiological features and the class vector. Then, to evaluate the emotion recognition, in the fully connected layer the cross-entropy loss function is used.

Each 1D CNN contains a sequence of temporal data for the recognition of local patterns, which can be learned from the physiological, video, and audio signals morphology. The functionality of the CNN layers is given by the convolution kernel that obtains the local patches and the Max pooling extracts the windows from the feature vectors to generate the down sampling output vector.

F. *Average Decision Fusion*

The output of the 4 fully connected layers of the three modalities are given as an input to an output generator with weighted decision fusion. The final output generator provided output by taking in emotion state produced by 4 fully connected layers and providing output by using average weighted decision fusion model. In this method, we take output of all 4 sub-models produced by the SoftMax function resulting in probabilities for all 10 emotion states. Suppose the 4 sub-models are represented by E_V, E_A, E_{EEG}, E_{E4}, Where E_V stands for output vector of Emotions produced by Visual modality, E_A stands for output vector of Emotion of Audio modality, E_{EEG} for output vector of EEG signal, and lastly E_{E4} gives the output vector produced by 6 bio-signals of Emperica E4 device.

Each output vector will contain probabilities of all 10 classes produced for that instance, say P_{v1}, P_{v2}, P_{v3}, P_{v4}, P_{v5}, P_{v6}, P_{v7}, P_{v8}, P_{v9}, P_{v10}, where P_{v1} represents, the probability of the output to be less cheerful and P_{v6} represents probability of belonging to very cheerful for visual modality model. This Probability value is generated by SoftMax activation function displayed in equation (2), employed after fully connected layers in each sub-model.

$$SoftMax\ (P_{Vi}) = \frac{\exp(P_{Vi}\ )}{\sum_i exp(P_{Vi}\ )} \tag{2}$$

Then in final output generator, probability of corresponding class will be added with probabilities of same class output from different sub-model as shown in equation (3).

$$P_i = P_{Vi} + P_{Ai} + P_{EEGi} + P_{E4i} \tag{3}$$

After calculating probability of each output class, the counter classes' probabilities are compared with each other, the one with more value will be provided as output, resulting in 5 outputs for a single instance.

$$E = max(P_{i,}P_{n-i}), i \in (1, \dots\dots, n) \tag{4}$$

From equation (4), it can be observed that for "n" output classes, there will be n/2 outputs.

For all the fully connected layers in four sub-models, Adam optimizer has been used, and loss is calculated by categorical cross-entropy as shown in equation (5).

$$Loss = -\sum\left(y'_{i1}log(y_{i1}) + y'_{i2}log(y_{i2}) + \cdots + y'_{in}log(y_{in})\right) \tag{5}$$

y_{i1}, y_{i2}, $y_{in}$are internal node labels, y'_{i1} , y'_{i2} , y'_{in} are the output layer nodes, produced by SoftMax function.

IV. Result & Discussion

The proposed model was executed twice once for output categorized as Dimensional Affect (low valance, high arousal, low arousal, high valance), and second execution for ten emotion states. For both executions, the proposed model works exactly same. The model was evaluated on AWS, Amazon EC2 G4dn instance having T4 GPU, with 100Gbps networking. The dimensional model presupposes that emotions are not independent and that there exists a relation between them hence the need to place them in a spatial space, whereas discrete emotion models categorize the emotions based upon the feelings of the user. Both have their own importance, to have an empathetic HCI, root should understand the actual discrete emotion, but at the same time should be able to understand the impact of the emotion using dimensional state. With rigorous experimentation, tuned hyper-parameter values were identified for most efficient results shown in table 1.

For performance evaluation the training and testing of the model is performed using "leave one out". The size of the dataset is limited, having all the signals recorded only for 10 minutes for each of the 32 participants, generating total of 320 minutes' data approximately. But only taken 18 participants data was taken for model execution, resulting in almost half the total data. Therefore, to have large training set, leave one out strategy was employed, implying that first model will be trained on 17 participants complete data, will be tested for performance evaluation of 18th participant. Then the same process will be repeated leaving another participant data for testing. So, the model will be executed 18 times for both types of output categories.

TABLE 1
Hyper-Parameters

	LEARNING RATE	EPOCHS	BATCH SIZE	DROPOUT	ACTIVATION	OUTPUT
VISUAL	0.001	60	32	0.5	ReLU	Soft Max
AUDIO		56	48	0.5		
EEG		120	32	-		
E4		68	16	-		

The execution of the model has been performed on AWS's Amazon EC2 instance through MacBook Pro. To evaluate the performance of proposed multimodal, F1-score has also been computed along with Accuracy.

TABLE 2
Model Performance for Emotion States

	P_1	**P_4**	**P_8**	**P_9**	**P_{10}**	**P_{11}**	**P_{15}**	**P_{18}**	**P_{19}**	**P_{20}**	**P_{21}**	**P_{22}**	**P_{23}**	**P_{24}**	**P_{25}**	**P_{26}**	**P_{29}**	**P_{30}**
									F1-SCORE									
LESS CHEERFUL	57.9	71.2	69.8	61.7	64.3	66.2	68.7	70.1	65.2	67.2	**72.3**	59.4	65.4	68.2	67.9	63.6	61.8	68.1
VERY CHEERFUL	70.5	69.3	72.3	**79.7**	76.1	73.5	68.5	69.3	70.4	61.9	78.4	73.7	62.7	69.5	69.8	66.3	70.3	75.8
LESS HAPPY	73.8	67.3	69.4	78.4	61.1	54.9	66.9	64.5	77.9	72.5	83.6	44.7	70.2	65.8	73.5	71.9	74.9	**79.1**
VERY HAPPY	81.9	74.9	78.5	80.3	72.8	75.1	**83.4**	73.9	71.8	72.9	51.4	78.2	68.9	72.3	70.3	68.4	63.7	75.3
LESS ANGRY	68.7	64.5	70.3	68.4	72.7	69.3	71.6	74.8	75.4	65.2	80.1	75.4	67.3	**82.5**	67.1	71.9	66.9	80.7
VERY ANGRY	75.4	58.4	68.4	67.6	78.3	72.6	**87.8**	51.9	70.3	81.9	71.8	71.7	76.4	72.1	74.9	72.3	71.5	57.9
LESS NERVOUS	56.2	65.8	71.2	75.4	74.1	79.4	75.5	66.3	73.9	72.6	79.5	72.8	79.8	77.9	64.8	80.1	78.4	**82.9**
VERY NERVOUS	69.2	72.5	75.3	76.1	63.8	68.4	**82.7**	78.2	67.3	69.8	64.9	81.6	54.9	76.8	75.8	73.6	40.6	79.4
LESS SAD	79.3	68.5	72.4	73.4	80.4	81.9	69.5	68.7	64.5	55.8	81.6	68.2	**82.3**	79.5	69.4	78.9	63.5	69.9
VERY SAD	74.6	**83.6**	76.7	72.5	75.8	78.4	78.2	66.1	62.6	79.3	76.7	79.6	74.8	73.1	77.2	82.4	79.9	73.6
									ACCURACY									
LESS CHEERFUL	68.4	63.5	75.3	71.2	69.9	68.1	64.5	74.7	69.9	64.3	70.4	66.7	71.7	73.2	71.9	66.8	**78.9**	68.7

VERY CHEERFUL	66.1	68.7	51.9	68.4	59.8	70.2	62.6	**76.1**	73.5	68.5	68.3	49.8	68.7	71.8	72.3	72.6	74.1	57.6
LESS HAPPY	70.8	71.8	66.3	70.2	75.3	65.7	68.4	68.1	72.7	70.8	66.1	72.5	75.1	63.8	64.5	67.6	**78.3**	72.6
VERY HAPPY	69.5	75.4	72.6	69.3	68.7	71.8	67.6	**78.3**	72.6	58.7	63.7	59.8	69.3	61.7	62.6	74.8	66.2	71.9
LESS ANGRY	64.5	75.1	68.5	70.3	76.1	69.3	72.3	74.7	75.1	62.5	66.9	69.3	**80.2**	66.2	63.6	61.8	68.9	72.3
VERY ANGRY	68.9	69.3	71.9	73.9	74.9	72.4	63.9	70.3	**79.4**	67.4	71.5	75.3	76.1	67.8	74.7	65.4	54.9	71.7
LESS NERVOUS	67.3	72.6	64.8	67.3	64.5	70.6	70.4	74.9	46.7	69.5	**81.7**	72.4	73.4	51.7	70	62.7	72.3	64.5
VERY NERVOUS	72.1	64.5	70.6	72.6	68.9	68.4	62.1	73.9	72.5	68.3	45.8	**76.9**	71.8	72.9	68.4	70.2	69.4	62.6
LESS SAD	**73.2**	62.6	68.4	68.1	67.3	66.9	54.7	67.3	59.4	65.4	68.2	76.1	73.5	68.5	44.3	68.9	63.6	61.8
VERY SAD	71.6	48.7	69.2	63.5	76.4	62.3	66.3	70.2	68.7	71.2	**73.8**	63.6	61.8	69.3	72.6	67.3	71.8	72.9

TABLE 3
MODEL PERFORMANCE FOR DIMENSIONAL AFFECTS

	P_{1}	P_{4}	P_{8}	P_{9}	P_{10}	P_{11}	P_{15}	P_{18}	P_{19}	P_{20}	P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}	P_{29}	P_{30}
F1-SCORE																		
LOW VALANCE	**89.6**	87.6	83.6	83.1	80.1	79.8	68.7	73.9	71.8	72.9	68.4	79.2	69.4	68.2	77.2	82.4	79.9	73.6
HIGH VALANCE	84.3	79.7	78.9	91.7	76.1	81.3	68.5	69.3	70.4	61.9	78.4	73.7	62.7	71.5	75.3	76.1	**92.6**	83.7
LOW AROUSAL	79.8	**91.7**	81.3	81.4	84.2	83.5	66.9	64.5	77.9	80.5	83.6	81.4	70.2	65.8	73.5	71.9	74.9	84.1
HIGH AROUSAL	83.2	78.9	78.5	80.3	**90.7**	86.7	83.4	73.9	71.8	72.9	80.6	78.2	68.9	72.3	70.3	71.4	81.5	75.3
ACCURACY																		
LOW VALANCE	75.4	58.4	68.4	72.3	74.7	75.1	82	57.8	77.2	**82.4**	79.9	73.6	76.4	72.1	74.9	72.3	71.5	68.3
HIGH VALANCE	56.2	65.8	71.2	**81.4**	73.7	79.4	75.5	66.3	73.9	72.6	79.5	72.8	79.8	77.9	64.8	80.1	78.4	78.1
LOW AROUSAL	69.2	72.5	75.3	77.8	79.7	**80.2**	82.7	78.2	67.3	77.2	79.4	79.9	73.6	76.8	77.2	78.1	76.9	70.6
HIGH AROUSAL	**79.8**	69.6	76.3	70.7	76.4	72.1	74.9	72.3	76.3	77.3	74.9	72.4	74.7	65.4	72.5	75.3	76.1	76.4

The model was evaluated on both accuracy as well as F1-score, as both are good evaluation metrics for machine learning models, although accuracy is more suitable for balanced dataset, while F1-score is suitable for both balanced as well as imbalanced dataset. The K-EmoCon is 50-50 balanced, but it doesn't fall in imbalanced dataset as well. We have employed both strategies because, some researchers only use F1-score, while others use accuracy. Authors expect the DREAM model to act as a baseline for future empathic HCI emotion recognition models, so we have provided both the evaluation metrics for the researchers. The model was executed twice to identify both types of emotion Dimensional as well as discrete, for the similar reason, so every researcher working on K-EmoCon, will be able to compare their work with DREAM results. Also authors have developed model for both types of emotions recognition, because identifying an emotion type depends upon the situation, for real-time application of empathic HCI emotion recognition, emotion should be both discrete and dimensional. Dimensional emotion gives us the amplitude of the emotion, whereas discrete emotion allows us to understand the actual emotion.

A. Result

The proposed work has three modalities, each executed separately in synchronous manner, combining the output of final layers of each model into Output layer. The performance evaluation of proposed model for both output categories are shown in table 2 and table 3. For emotional states highest F1-score was obtained for very angry emotion for P_{15}, and highest accuracy of 81.7 for P_{21} participant for Less Nervous emotion. For Dimensional Affect Arousal and Valance, the highest accuracy of 82.4 and F1-score of 92.6 was achieved. DREAM model have performed better than the SOTA which was achieved earlier by Alhussein et. al. [34], where they have achieved 82% and 83% F1-score for arousal and valance emotion respectively. Alhussein has not performed the discrete emotion recognition, so we are only able to compare the model on dimensional emotional, where DREAM has outperformed the SOTA model [34].

B. Discussion

The proposed multimodal have resulted in state-of-the-art results for both emotion categories. The comparison of the model is not possible, since this is the first work that incorporate three different modalities on K-EmoCon dataset and for classifying that many different emotions. Although the model proposed by Alhussein et al. [34] have produced the earlier state of the art on dimensional emotions, but our model has surpassed their model performances.

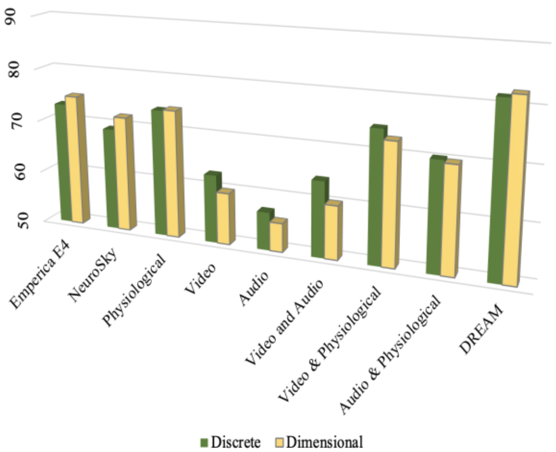

Fig.4. Ablation Study Analysis

The ablation study on the proposed model was performed as well, to analyze the impact of each modality. The Physiological modality performs best individually, followed by video modality, and the accuracy achieved by only speech signals for emotion detection was the least of all three. The best accuracy achieved by each modality in the ablation study is shown in figure 4. The variation in results in each individual modality is impacted by the amount of the data available as well, and the bio-signals were available for more participants in comparison to video or audio signals. For future work, researchers should work on making the model more secure and fast, rather than just enhancing the performance of the models.

V. Conclusion

Human computer interaction using consumer grade camera and sensors can enable the use of robots for various domains like, healthcare, hotel management, smart industries, supply chain manager, Human resources, and many more. With availability of smart wearable sensors, it is possible to track the activities of a person along with their audio, video, and bio-signals. We proposed multimodal emotion recognition using K-EmoCon dataset, containing audio, video, and physiological signals. No other work has used all the modalities for emotion recognition. The performance of the model was also tested on two types of emotion classification. In both classifications, DREAM has produced state of the art results. For improving the model further researchers can embed other modalities like linguistic by generating the transcripts of audio signals, from video signals, researchers can also use hand movements (gestures), for more accurate emotion recognition.

References

[1] D. J. Cook, J. C. Augusto, & V. R. Jakkula, “Ambient intelligence: Technologies, applications, and opportunities”. Pervasive and Mobile Computing, 5(4), 277-298, 2009.

[2] D. Ayata, Y. Yaslan, Y., & M. E. Kamasak, (2018). Emotion based music recommendation system using wearable physiological sensors. IEEE transactions on consumer electronics, 64(2), 196-203.

[3] J. Ploennigs, J. Cohn, & A. Stanford-Clark, “The future of IoT”. IEEE Internet of Things Magazine, 1(1), 28-33, 2018.

[4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, & M. Zorzi, “Internet of things for smart cities”. IEEE Internet of Things journal, 1(1), 22-32, 2014.

[5] Fei Yan, Abdullah M. Iliyasu, and Kaoru Hirota. "Conceptual framework for quantum affective computing and its use in fusion of multi-robot emotions." Electronics 10.2 (2021): 100.

[6] Rui, Yong. "From artificial intelligence to augmented intelligence." IEEE MultiMedia 24.1 (2017): 4-5.

[7] M. Obrist, C. Velasco, N. Ranasinghe, A. Israr, A. Cheok, & K. P. Gopala, “Sensing the future of HCI: touch, taste, and smell user interfaces. interactions,” 23(5), 40-49, 2016.

[8] M. Chowdary, Tu Kalpana, N. Nguyen, and D. Jude Hemanth. "Deep learning-based facial emotion recognition for human–computer interaction applications." Neural Computing and Applications, 1-18, 2021.

[9] P. Gupta, S. A. Balaji, S. Jain, and R. K. Yadav. "Emotion Recognition During Social Interactions Using Peripheral Physiological Signals." In Computer Networks and Inventive Communication Technologies, pp. 99-112. Springer, Singapore, 2022.

[10] J. S. Park, J. H. Kim, & Y. H. Oh, (2009). Feature vector classification based speech emotion recognition for service robots. IEEE Transactions on Consumer Electronics, 55(3), 1590-1596.

[11] A. Guhn, L. Merkel, L. Hübner, I. Dziobek, P. Sterzer, & S. Köhler, “Understanding versus feeling the emotions of others: how persistent and recurrent depression affect empathy”. Journal of psychiatric research, 130, 120-127, 2020.

[12] S. Saha, S. Datta, A. Konar and R. Janarthanan, "A study on emotion recognition from body gestures using Kinect sensor," 2014 International Conference on Communication and Signal Processing, 2014, pp. 056-060

[13] Jie Wei, Xinyu Yang, and Yizhuo Dong. "Time-Dependent Body Gesture Representation for Video Emotion Recognition." In International Conference on Multimedia Modeling, pp. 403-416. Springer, Cham, 2021

[14] D. Bertero and P. Fung, "A first look into a Convolutional Neural Network for speech emotion detection," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 5115-5119.

[15] A. Kumar, K. Sharma, and A. Sharma, "MEmoR: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries." Image and Vision Computing (2022): 104483.

[16] A. Sharma, K. Sharma, and A. Kumar, "Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion." Neural Computing and Applications(2022): 1-14.

[17] P. S. Gupta, A. Balaji, S. Jain, and R. K. Yadav, "Emotion Recognition During Social Interactions Using Peripheral Physiological Signals." In Computer Networks and Inventive Communication Technologies, pp. 99-112. Springer, Singapore, 2022.

[18] J. Quan, Y. Miyake, and T. Nozawa, "Incorporating Interpersonal Synchronization Features for Automatic Emotion Recognition from Visual and Audio Data during Communication." Sensors 21, no. 16 (2021): 5317.

[19] S. Lalitha, S. Tripathi, & D. Gupta, “Enhanced speech emotion detection using deep neural networks”. International Journal of Speech Technology 22, 497–510 (2019).

[20] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "LSSED: a large-scale dataset and benchmark for speech emotion recognition." In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 641-645. IEEE, 2021.

[21] M. Chowdary, Tu Kalpana, N. Nguyen, and D. Jude Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications." Neural Computing and Applications (2021): 1-18.

[22] P. Gupta, S. A. Balaji, S. Jain, and R. K. Yadav. "Emotion Recognition During Social Interactions Using Peripheral Physiological Signals." In Computer Networks and Inventive Communication Technologies, pp. 99-112. Springer, Singapore, 2022.

[23] M. Zitouni, Cheul Sami, Young Park, Lee Uichin, Leontios Hadjileontiadis, and Ahsan Khandoker. "Arousal-Valence Classification from Peripheral Physiological Signals Using Long Short-Term Memory Networks." In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 686-689. IEEE, 2021.

[24] F. A. Alskafi, A. H. Khandoker, and H. F. Jelinek. "A Comparative Study of Arousal and Valence Dimensional Variations for Emotion Recognition

Using Peripheral Physiological Signals Acquired from Wearable Sensors." In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1104-1107. IEEE, 2021.

[25] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara. "SigRep: Toward Robust Wearable Emotion Recognition With Contrastive Representation Learning." IEEE Access 10 (2022): 18105-18120.

[26] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations." Scientific Data 7.1 (2020): 1-16.

[27] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[28] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang. "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6124-6128. IEEE, 2020.

[29] A. Kumar, K. Sharma, and A. Sharma, "Hierarchical deep neural network for mental stress state detection using IoT based biomarkers." Pattern Recognition Letters 145 (2021): 81-87.

[30] M. Z. Uddin, J. J. Lee, and T-S. Kim. "An enhanced independent component-based human facial expression recognition from video." *IEEE Transactions on Consumer Electronics* 55.4 (2009): 2216-2224.

[31] U. Ahmed, J. C. W. Lin, and G. Srivastava. "Emotional intelligence attention unsupervised learning using lexicon analysis for irony based advertising." *ACM Transactions on Asian and Low-Resource Language Information Processing* (2022).

[32] Y. Zhang, G. Srivastava, “Speech emotion recognition method in educational scene based on machine learning”, *EAI Endorsed Transactions on Scalable Information Systems, EAI, 2022*. DOI: 10.4108/eai.10-2-2022.173380.

[33] K. Yang, B. Tag, Y. Gu, C. Wang, T. Dingler, G> Wadley, & J. Goncalves, “Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer”. In Proceedings of the 2022 International Conference on Multimedia Retrieval (pp. 562-570).

[34] G. Alhussein, M. Alkhodari, A. Khandoker, & L. Hadjileontiadis, “Novel Speech-Based Emotion Climate Recognition in Peers’ Conversations Incorporating Affect Dynamics and Temporal Convolutional Neural Networks”, 2023.

[35] A. Kumar, K. Sharma, & A. Sharma, A. “Empirical Analysis of Psychological Well-Being of Students During the Pandemic with Rebooted Remote Learning Mode”. In Proceedings of Data Analytics and Management: ICDAM 2022 (pp. 13-29). Singapore: Springer Nature Singapore, 2023.

Dr. Aditi Sharma (Member, IEEE) is working as a Regular Assistant Professor at Thapar Institute of Engineering and Technology. She has near 5 years of experience in Academia. She has previously worked with Jaypee University of Information Technology, Waknaghat as Assistant Professor (Grade- II) in Department of Computer Science and Engineering/Information Technology. Prior to Jaypee Group, Ms. Sharma was associated with Delhi Technological University, and DIT University.

Dr. Aditi has received her Ph.D. from Delhi Technological University in the field of Affective Computing. She earned her Post-graduation, Masters in Software Engineering from Delhi Technological University in 2017. She graduated from Punjabi University Patiala in 2015 with Bachelor of Technology in Computer Science and Engineering. With a top score of 717 and an All India Rank of 571 in 2015, she has qualified GATE four times. In 2018, She also qualified UGC-NET for Assistant Professor. Affective Computing, Machine Learning, Predictive Healthcare, and Text Summarization are some of her research interests. She has 14 publications, five of which are in high impact SCI/SCIE journals. In 2022 and 2023, Dr. Sharma has received “Commendable Research Award for Excellence in Research” for her research work from Delhi Technological University. She holds professional memberships in IEEE, ACM, CSI and IAENG.

Dr. Akshi Kumar (Senior Member, IEEE) is a Senior Lecturer, and Director-PGR in the Department of Computing at the Goldsmiths, University of London, London, United Kingdom. She is a Post-doc from Federal Institute of Education, Science and Technology of Ceará, Fortaleza, Brazil and a PhD from Faculty of Technology, University of Delhi, India. She has worked as an Associate Professor at the Manchester Metropolitan University, Manchester, United Kingdom and Netaji Subhas University of Technology (NSUT), New Delhi, India and as an Assistant Professor at the Delhi Technological University (DTU), New Delhi, India.

Dr. Kumar has been endorsed by the Royal Academy of Engineering, United Kingdom as an Exceptional promise in the field of Artificial Intelligence/Data Science in 2022. She has received 9 research awards for Excellence in Research from various National and International organizations. Her name has been included in the *“Top 2% scientist of the world”* list by Stanford University, USA in 2023, 2022 and 2021. Based on the list, her current world rank within the field of Artificial Intelligence is 1415. She has published more than 95 peer-reviewed journal papers including 55 SCIE publications, 65+ conference papers with 5 best paper awards and 2 patents with Indian Patent Office. She has successfully guided 6 doctorates, 33 Master thesis candidates. She has been serving as an Associate editor and guest editor in various high impact journals with reputed publishers. Her research interests are in affective computing, social network and media analytics, NLP, and AI for pervasive healthcare.