Spontaneous Facial Behavior Analysis using Deep Transformer Based Framework for Child–Computer Interaction

Article *in* ACM Transactions on Multimedia Computing, Communications and Applications · May 2022

DOI: 10.1145/3539577

CITATIONS
2

READS
60

4 authors, including:

Abdul Qayyum
University of Burgundy
111 PUBLICATIONS **1,169** CITATIONS
SEE PROFILE

M. Tanveer
Indian Institute of Technology Indore
197 PUBLICATIONS **4,857** CITATIONS
SEE PROFILE

Moona Mazher
University College London
54 PUBLICATIONS **274** CITATIONS
SEE PROFILE

Spontaneous Facial Behavior Analysis using Deep Transformer Based Framework for Child–Computer Interaction

ABDUL QAYYUM, Department of Computer Science and Engineering, Université de Bourgogne, France,
IMRAN RAZZAK, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
M. TANVEER, Department of Mathematics, Indian Institute of Technology Indore, Simrol, Indore, 453552, India
MOONA MAZHER, Department of Computer and Mathematics, University Rovira i Virgili, Tarragona, Spain

Abstract: A fascinating challenge in robotics-human interaction is imitating the emotion recognition capability of humans to robots with the aim to make human-robotics interaction natural, genuine and intuitive. To achieve the natural interaction in affective robots, human-machine interfaces, and autonomous vehicles, understanding our attitudes and opinions is very important, and it provides a practical and feasible path to realize the connection between machine and human. Multimodal interface that includes voice along with facial expression can manifest a large range of nuanced emotions compared to purely textual interfaces and provide a great value to improve the intelligence level of effective communication. Interfaces that fail to manifest or ignore user emotions may significantly impact the performance and risk being perceived as cold, socially inept, untrustworthy, and incompetent. To equip a child well for life, we need to help our children identify their feelings, manage them well, and express their needs in healthy, respectful, and direct ways. Early identification of emotional deficits can help to prevent low social functioning in children. In this work, we analyzed the child's spontaneous behavior using multimodal facial expression and voice signal presenting multimodal transformer-based last feature fusion for facial behavior analysis in children to extract contextualized representations from RGB video sequence and Hematoxylin and eosin video sequence and then using these representations followed by pairwise concatenations of contextualized representations using cross-feature fusion technique to predict users emotions. To validate the performance of the proposed framework, we have performed experiments with the different pairwise concatenations of contextualized representations that showed significantly better performance than state of the art method. Besides, we perform t-distributed stochastic neighbor embedding visualization to visualize the discriminative feature in lower dimension space and probability density estimation to visualize the prediction capability of our proposed model.

CCS Concepts: • **Human-centered computing** → **Interaction design theory, concepts and paradigms**; *Interaction design*; • **Machine learning** → *Machine learning algorithms.*

Additional Key Words and Phrases: Datasets, neural networks, gaze detection, text tagging

Authors' addresses: Abdul Qayyum, eng@corporation.com, Department of Computer Science and Engineering, Université de Bourgogne, France, P.O. Box 1212, Dijon, France, ; Imran Razzak, School of Computer Science and Engineering, University of New South Wales, Sydney, 1 Thørväld Circle, Australia, imran.razzak@unsw.edu.au; M. Tanveer, Department of Mathematics, Indian Institute of Technology Indore, Simrol, Indore, 453552, Indore, India; Moona Mazher, Department of Computer and Mathematics, University Rovira i Virgili, Tarragona, Spain.

1551-6857/2022/5-ART $15.00
https://doi.org/10.1145/3539577

1 INTRODUCTION

In our daily interactions with people around us, emotional response (emotions/reaction) play an essential role in efficient communication and influence all involved in the conversation. Emotions pervade each aspect of interaction so deeply that we often notice extreme emotions such as upset, anger, etc. However, the impact of emotions is far more than this, and it impacts our perception, attention, memory, and decision-making capabilities. Traditionally, human-computer interaction is considered the "ultimate" exception. A user must ignore his emotional selves to interact rationality and efficiently with the machine. If we consider a computer or an interface merely as a tool, then designing an interface should consider application-specific goals evaluated through metrics as learnability, usability, efficiency, and accuracy. Nevertheless, if we consider the user interface as a medium between us and the machine, then it is important to think about usability, learnability, and gratifications. Recent research in user-interface design and psychology recommended different views of the relationship between humans, computers, and emotion. In the last decade, the psychology of emotion has been explored significantly [9]. It may be due to technological advancement, i.e., handheld devices that eased and provided different methods for machine-human interaction. The emotion recognition market was valued at USD 19.87 million in 2020, lessons which is expected to triple by 2026, registering a CAGR of 18.01% during the forecast period (2021-2026). Improvements in machine learning, signal processing, advancement, and cheap hardware enabled us to analyze physiological correlations of emotions, i.e., even our personal computer can make a judgment on our emotional state [4, 9, 17]. 'The source code is publicaly avaialble at[1]

Age/Emotions	Happy	Sad	Angry	Surprised	Disgusted	Fearful
6 Years	83%	76%	70%	57%	30%	40%
7 Years	91%	85%	78%	49%	25%	50%
8 Years	93%	72%	69%	77%	39%	47%
9 Years	96%	76%	76%	70%	40%	53%
10 Years	98%	77%	71%	83%	42%	50%
11 Years	82%	73%	72%	80%	53%	66%
12 Year	96%	78%	67%	81%	62%	55%
13 Years	97%	78%	73%	86%	61%	57%
14 Years	99%	76%	79%	91%	75%	60%

Fig. 1. Spontaneous facial expressions

Understanding the emotions holds significance during the interaction between our and machine communication systems. Emotion recognition improves human and computer interfaces and enhances the feedback mechanism actions taken by computers from the users. With the development of smart and intelligent solutions such as smart homes or personal health, emotions may play a significant role in human-machine interaction, and its market is expected to reach USD 52.86 million by 2026. A multimodal user interface that includes voices, faces, and bodies can manifest a large range of our emotions compared to earlier text-based interfaces [24]. A user interface that ignores emotions or does not manifest appropriate emotions may significantly impact the performance and add an additional risk of being perceived as socially inept, incompetent, and untrustworthy. Both adults

[1]https:github.comRespectKnowledgeChild_Video_FacialExpression_Deep-Learning

Fig. 2. Spontaneous facial expressions

and children are equally impacted where the contribution of emotional competence to social competence has long-term implications. Emotions are essential for communication as well as for social interaction and play a crucial role in decision-making hence critical in our daily lives, i.e., how we engage with others and live our lives, how we feel others' emotions, etc. Facial expression and speech signals are considered the most effective means to convey our feelings.

Facial expression is one of the most effective approaches to delivering information on our emotional feelings to others. Understanding emotional skills are essential for healthy social behavior development, especially in childhood. These emotional skills are related to very important development outcomes, i.e., mathematics skills, school readiness, enhanced literacy and language. Emotion recognition plays a very important role in designing a practical and learnable interface for efficient communication. With the advancement in autonomous devices, facial expression recognition becomes very important. A fascinating challenge in robotics-human interaction is imitating the emotion recognition capability of humans to robots with the aim to make human-robotics interaction natural, genuine and intuitive. To achieve the natural interaction in affective robots, human-machine interfaces, and autonomous vehicles, understanding our attitudes and opinions is very important, and it provides a practical and feasible path to realize the connection between machine and human.

Despite early discrimination of emotional expressions, the ability to accurately label emotions continues to develop throughout childhood and adolescence, recognizing some types of emotions developing earlier than others. A study conducted on children of age 5, 7, 9, and 11 year-old children showed that emotion recognition skills improve with age, although specific types of faces have may different recognition than others, i.e., children younger than five years older are even able to recognize sad and happy emotions with the same accuracy of an adult, however, they develop skills for other emotions slowly. Understanding children's emotions at an early age may help to analyze deficits in emotion recognition; hence we can overcome social functioning through the user interface. Existing methods mainly focus on posed facial expressions in adults; however, our expressions are naturally different from posed expressions in the real world. Besides, recognition of emotion in children, especially spontaneous, is considered less comparatively. Analysis of spontaneous facial expressions also helps to deal with deficits in children by developing an efficient user interface. To deal with the challenge mentioned above, in this work, we explored the problem of spontaneous emotion analysis in children and proposed a novel end-to-end framework consisting of a lightweight transformer-based network. We present different multimodal pre-trained

transformers-based framework. We believe that the contextualized representations extracted from RGB video and Hematoxylin and eosin video sequences capture the important information, improving the performance of the downstream emotion recognition task. The key contributions of this work can be described as

- We present transformers-based multimodal architecture to extract contextualized representations from RGB video sequence and Hematoxylin and eosin video sequence and then use these representations to predict user's emotions.
- We present the late fusion of RGB video and Hematoxylin and eosin video through pairwise concatenations of contextualized representations using the cross-feature fusion technique, which results in robust and efficient contextualized representations.
- We have used different pre-trained transformers and performed the late feature fusion of the RGB video sequence and Hematoxylin and eosin video sequence. We further validate the robustness of the proposed framework through t-distributed stochastic neighbor embedding visualization to visualize the discriminative feature in lower dimension space and probability density estimation to visualize the prediction capability of our proposed model.
- We have conducted extensive experiments on a benchmark child facial expression dataset (spontaneous child-computer interaction) that showed significantly better performance than the state-of-the-art methods.

2 RELATED WORK

A few years back, when we talked about giving emotions to the computer, it looked like fiction movies, the computer is a self-aware intelligent machine that can feel feelings. The activities, tasks and domains the computer is performing and the way we use it, are continuously evolving. With the development of smart devices, an additional dimension has been added to machine-human interaction, which considers the utility and effectiveness of emotions in human-machine interaction by incorporating emotions in technologies. It looks like these systems will take over the world by engaging us in almost all activities naturally. Around two decades back, R. Picard coined the term "affective computing" for computing that relates to, arises from, or deliberately influences the emotional state or other affective phenomena1 in human-machine interaction. Since then, several researchers have focused on analyzing the impact of emotions on human-computer interaction. The development of smart and intelligent solutions such as smart homes or personal health, makes emotions a crucial component in human-machine interaction.

Interactive, handheld, and inexpensive technology enabled us to analyze the physiological correlation of our emotions. The dramatic improvement in machine learning, quality of signal processing, and hardware allowed us to analyze our judgment about our emotions and their role in decision making. Emotions play a critical role in our daily lives, how we engage with each other in our daily lives, and how it affects our decision-making. In 4 B.C., Aristotle was the first to identify the exact number of core human emotions known as Aristotle's List of Emotions such as kindness, confidence, fear, anger, friendship, envy, calm, enmity, shame, pity, indignation, shamelessness, emulation, and contempt. In 1970, Paul Eckman identified six basic human emotions such as surprise, anger, fear, sadness, happiness, and disgust that are universally experienced worldwide, which were extended to a few more embarrassment, shame, pride, and excitement later on. With the rapid advancement in handheld devices and their increasing usage in society, the need for technology to assess potential customers and find an appropriate alternate is increasing dramatically. Emotion recognition through facial expression is one of the most powerful, natural, and universal approaches to conveying our emotional states and intentions. The majority of the traditional facial expression methods are based on handcrafted features or shallow learning, such as LPP [13] non-negative matrix factorization [27] local binary patterns (LBP) [16], LBP on three orthogonal planes [26] and sparse learning [28], CNN [8, 19, 20, 22, 25]. The success of deep learning in computer vision tasks has also been applied to emotional

recognition. Yao et.al presented HoloNet for emotion recognition in Wild (EmotiW) 2016 challenge [21]. Unlike simple and shallow networks, HoloNet enhances the non-standard non-linearity in earlier convolutional layers and reduces the redundant features using rectified linear units. To construct the middle layers, they combine residual structure and concatenated rectified linear unit and broaden the network; the last layers are designed as a variant of the inception network. Thus HoloNet is able to capture multi-scale features, explicitly the emotions. Large data is one of the major challenges in facial expression recognition. Multiple datasets can be combined to increase the dataset size; however, inconsistent annotations in facial expression datasets are inevitable and can have a huge impact on the performance. Zheng et. al. presented inconsistent pseudo annotations to latent truth frameworks to deal with the problem of dataset inconsistency using a trainable LTNet that finds the latent truths from machine annotations and human annotations by maximizing the log-likelihood on inconsistent annotations [23]. Wang et.al; focused on real-world pose and occlusion in order to improve the recognition of facial expression using a region attention network (RAN) that captures the importance of different regions adaptively for occlusion and pose variant recognition [18]. Region attention network aggregates and embeds various facial region features into compact fixed-length representation, and finally, region-biased loss encourages high attention weights for important facial regions. Moao et al. presented CNN and MobileNet based real-time framework for facial expression recognition [14]. Experiments were conducted on JAFFE and CK+ dataset that showed 95.24% and 96.92% accuracy for emotion detection. Demisse et al. presented deformation-based representation to analyze 3D facial expressions [5]. Group structure is used to decouple the neutral face followed by non-linear facial expression manifold captured through mapping to linear space. Lue et al. presented expression-aware emotional color transfer framework to overcome the ambiguity between the emotions [10]. The approach consist of two phases, first emotions are predicted using classification followed by pre-trained color transfer model. Finally, emotional model is matched for color tranfer to target image.

Mehdizadehfar et al. analyzed the emotions of fathers of autistic children using EEG signals to understand whether the fathers of children with autism conditions have problems in labeling three facial expressions, including anger, sadness, and happiness [12]. AlBraikan et al. developed mobile framework that uses wearable sensor data to recognition five basic emotions to overcome the infirmity[1]. In another work, AlBraikan et al. presented hybrid sensor based fusion on stacking model which allows data from multiple source and developed an emotion framework to jointly embedded within a user-independent model [3]. Similarly, AlBraikan has deployed sequential model-based optimization for segmentation and feature selection to predict emotions form multimodal sensors [2]. Tobon et al. presented comprehensive review for EEG based analysis of emotional conditions [**?**]. Wavelet power, power spectra, and group independent component analysis were used to analyze the EEG signals. In another work, Mayor-Torres et al. evaluated the interpretability of deep learning methods for emotion recognition using EEG [11]. Evaluation of CNN for the emotion recognition using EEG signals is performed with the RemOve-And-Retrain approach for the recovery of emotion-relevant features and compared the performance with PatternNet, Pattern-Attribution, Layer-Wise Relevance Propagation (LRP), and Smooth-Grad Squared. Gu et al. presented EEG based study to analyze facial emotion recognition ability in hearing controls and deaf children [6]. Experiments suggest that deaf children showed lower performance in comparison to hearing controls. With the recent success of transformers in natural language processing, researchers focused on development of transformers to replace with traditional CNN architecture for vision applications. Pre-training transformer with self-supervised learning may help to overcome the lack of labeled data. Thus, in this work, we present pre-trained late fusion of transformers on RGB video and Hematoxylin and eosin video through pairwise concatenations of contextualized representations using the cross-feature fusion technique.

Fig. 3. Childern emotions with emotions and fear

3 EMOTION RECOGNITION IN CHILDREN

Proliferation of robotics made our life convenient and easy. One of the key weakness in robotics-human interaction is lack emotional understanding in machines which limit it capability to interact with us naturally. Imitating the human emotion recognition capability to robots to make human-robotics interaction natural, genuine and intuitive. To achieve the natural interaction in affective robots, human-machine interfaces and autonomous vehicles, understanding our attitudes and opinions is very important and it provides a practical and feasible path to realize the connection between machine and human. Facial expression recognition is widely explored topic however, most of the research work focus on posed expression, hence can not be applied directly in natural environment as real-world scenarios are spontaneous expression. Furthermore, spontaneous facial expression analysis in children is very less explored despite of its important i.e. early detection of facial expression deficit may help to prevent functioning in later age. The expression recognition capability continuous to develop throughout childhood and adolescence. To really equip a child well for life, we need to help our children to identify their feelings, manage them well, and express their needs in ways that are healthy, respectful and direct. Early identification of emotions deficits can help to prevent the low social functioning in children. In this work, we analyzed the child spontaneous behavior using multimodal facial expression and voice signals. We present multi-modal transformer based last feature fusion for facial behavior analysis in children. Figure **??** shows the framework architecture. First, we have obtained the contextualized representations from video sequences, the we

have preformed late fusion of representation obtained from RGB video sequence and Hematoxylin and eosin video sequence transformers.

Fig. 4. Six common facial emotion

For the first step, we have used pre-trained (ImageNet) encoder-based deep neural network with an attention mechanism to extract the contextualized representations from both the RGB video sequence and Hematoxylin and eosin video sequence. The main component of the video encoder is the transformer pre-trained on ImageNet. We then performed a pair-wise late fusion of both transformers then fine-tuned the framework for emotion recognition tasks using labeled video data. The heart of our framework is a transformer encoder that learns the contextualized representation from video sequences. The transformer provides the contextual information using an attention mechanism with an attention function considered mapping of the query and a group of key-value pairs to output. The position of each output pays attention to all inputs. We have used several attention mechanisms to create various representations of video signals. The encoder is constructed by stacking several layers consisting of a multihead attention module, followed by a fully connected layer. To process the video sequences (RGB video sequence and Hematoxylin and eosin video sequence), we first encode the feature vectors of dimension d models.

$$Z_{lp} = MSA\Big(NormL(Z_{(l-1)}\Big) + Z_{(l-1)}, \quad wherel = 1, ..., L \tag{1}$$

Where MSA is the multihead self-attention block, $Z_{(l-1)}$ is the previous layers before the MSA block. Z_{lp} is the layer block after MSA. The residual block used to add the layer block ($Z_{(l-1)}$) with MSA block. NormL is the normalization layer used before the MSA block.

$$Z_{fc} = MLP\Big(NormL(Z_{lp})\Big) + Z_{lp}, \quad l = 1, ..., L \tag{2}$$

The fully connected layer-1 (Fc1) with features size (3072x768), ReLU activation (ReLU), dropout layer for regularization, and fully connected layer-2 (Fc2) with feature size (768x5) are used at the end of both pre-trained transformers. The F_1 is the feature extraction from the last layer of the pre-trained vision transformer(vit_base_patch16_224). The F_2 is the features are extracted from the last layer of vision transformer(vit_base_patch16_224_miil).

$$F_1 = MLP_1\Big(NormL(Z_{lp1})\Big) + Z_{lp1}, \quad l = 1, ..., L \tag{3}$$

$$F_2 = MLP_1\Big(NormL(Z_{lp2})\Big) + Z_{lp2}, \quad l = 1, ..., L \tag{4}$$

We present the late fusion of RGB video and Hematoxylin and eosin video through pairwise concatenations of contextualized representations using the cross-feature fusion technique, which results in robust and efficient contextualized representations.

$$F_1F_2 = concat[MLP_1\Big(NormL(Z_{lp1})\Big) + Z_{lp1}, MLP_2\Big(NormL(Z_{lp2})\Big) + Z_{lp2}] \tag{5}$$

$$F_2F_1 = concat[MLP_2\Big(NormL(Z_{lp2})\Big) + Z_{lp2}, MLP_1\Big(NormL(Z_{lp1})\Big) + Z_{lp1}] \tag{6}$$

Finally, we get

$$Pff = concat[F_1F_2, F_2F_1] \tag{7}$$

where Pff is the pairwise feature fusion that concatenated the both contextual representation.. The feature dimension concatenated at each level is shown in Figure.1.

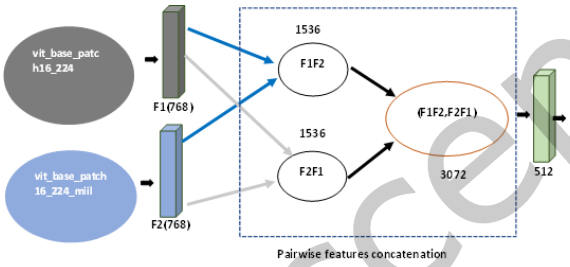


Fig. 5. Pairwise feature concatenation used using multi-model vision transformer for child facial expression detection

Keep in mind that the original vision-based transforms are trained on 1000 ImageNet dataset classes. We have changed the last layers to finetuned pre-trained transformers according to the number of classes. In our cases, we considered the five classes of emotions (fear, disgust, happy, sad, and surprise). Hematoxylin and eosin (H&E) stains have been used for at least a century for cancer diagnosis because it can essentially differentiate cytoplasmic, nuclei, and extracellular matrix features. According to the principle of H&E stain, hematoxylin stains cell nuclei blue, and eosin stains the extracellular matrix and cytoplasm pink. Intending to use the above unique characteristic of H&E stain as Hematoxylin-aware guidance for the facial expression classification task, we have used a transformer-based approach on H&E stain video sequences. To achieve H&E transformer, we apply a color decomposition technique to decompose the Hematoxylin Component from the original RGB image. This approach is commonly utilized as a color normalization preprocessing in traditional methods due to its robustness of color inconsistency in the H&E stained WSI. So each specific stain can be characterized by a specific optical density vector. The hematoxylin resulted in o values [0.18,0.20, 0.08] for R, G, and B channels. We can map RGB color space to any strain-specific color space with this color representation model. We extracted the Hematoxylin component based on this model by applying the color deconvolution method proposed. To pre-train our encoder, we employed ImageNet, followed by fine-tuning on RGB video and Hematoxylin and eosin video sequences. We fine-tuned our model on six basic emotion classes. We believe that the contextualized representations extracted from RGB video and Hematoxylin and eosin video sequences capture the important information, which improves the performance of the downstream emotion recognition task. We have trianed different transformers in parallel to extract the contextualized representation and performed pairwise late fusion of extracted representation.

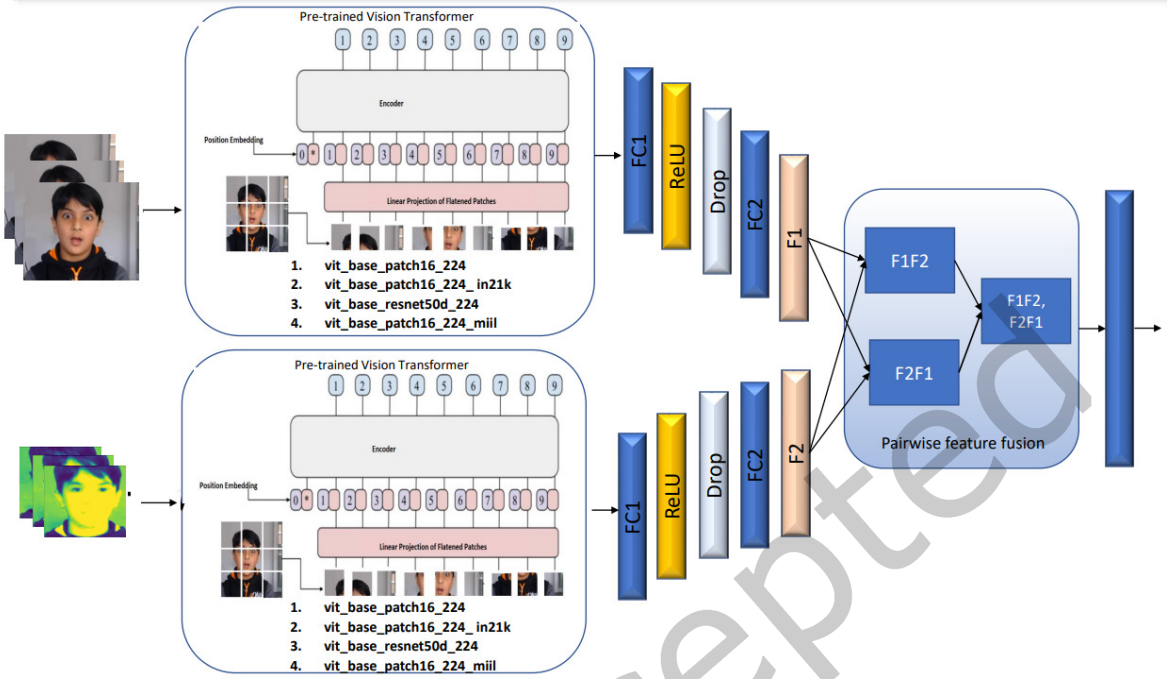


Fig. 6. Architectures of proposed Multimodal pairwise vision transformers

4 EXPERIMENT

In this section, we present experimental choices taken to evaluate our approach for a downstream task of facial emotion recognition on video sequences. We first present dataset, network parameter optimization, experimental analysis followed by benchmark evaluation. We have performed 10-fold validation on child spontaneous facial expression dataset. We have performed experiments with different models and performed late fusion to achieve better performance using different evaluation measures such as precision, recall, F-score, specificity, sensitivity, specificity and area under the curve (AUC). The Pytorch library is used for model development, training, optimization, and testing and also other libraries based on python are used for pre-processing and analysis of the datasets. The NumPy used to process the input array. The detail of environment with requirements that is necessary step for sitting the software and hardware used in training, optimization and validation of our proposed model is shown in Table.

4.1 Dataset

To really equip a child well for life, we need to help our children to identify their feelings, manage them well, and express their needs in ways that are healthy, respectful and direct. Early identification of emotions deficits can help to prevent the low social functioning in children. Most of the recent research focus on development of posed emotion recognition dataset. One of the recent effort by Khan et al. is development of children's spontaneous emotional database (LIRIS-CSE). The LIRIS-CSE dataset is spontanieos facial expression videos collected from different ethnicity's children. The dataset consist unconstrained videos with six basic emotions "surprise", "sadness", "happiness", "disgust" "neutral",and "fear". The dataset consist of 26,000 frames labelled by 22 experts, collected in unconstrained environment.

Table 1. NMVT Parameter description

Parameters	Values
Batch size	24
Total epochs	50
Optimizer	Adam
Initial learning rate	0.0001
Stopping criteria, and optimal model selection criteria	Stopping criterion is reaching the maximum number of epoch (5).
Training time	40 mints

Table 2. Evaluation of Proposed Lightweight ShallowNet (%)

	support	precision	recall	f1-score
fear	529	0.998	0.987	0.998
disgust	97	1.00	1.00	1.00
happy	778	1.00	0.998	0.987
sad	512	0.998	1.00	1.00
surprise	546	0.996	1.00	0.982
Macro avg.	2462	0.998	0.998	0.987
Weighted avg.	246	**0.9986**	**0.9989**	**0.9987**

4.2 Network Parameters

The proposed multimodal pre-trained transformers based architecture consist of transformer trained on RGB video sequences and Hematoxylin and eosin video sequence. There are different parameters that are required to be optimal. We set the learning rate to 0.0001 with Adam optimizer. As a loss function between the output of the model and the ground-truth sample, we have used weighted cross-entropy function. As the dataset is highly imbalance, thus we have used inverse class frequencies for weight balancing has been used to calculate the weighted cross entropy loss function. The higher-class samples require less weight and less class samples needs more weight values. The 24 batch-size with 50 number of epochs has been used with 5 early stopping steps. The best model weights have been saved for prediction in the validation phase. The input sequence size (1x2500) was used for training and prediction. The V100 tesla NVidia-GPU machine is used for training and testing the proposed model. The dataset is normalized between 0 and 1 using the max and min intensity normalization method. The detail of training protocol is shown in Table 1.

4.3 Evaluation Metrics

In order to evaluate the classification performance of proposed ShallowNet, we have used different evaluation metrics as precision, recall, F-score, specificity, sensitivity, and area under curve (AUC). Precision is referred as positive predictive value (PPV) and is the true positive relevant measure calculated as, $P = \frac{TP}{TP+FP}$. Recall is referred to as the true positive rate or sensitivity, is the ratio of correctly predicted positive observations to all observations in the actual class. Recall is calculated as $R = \frac{t_p}{t_p+f_n}$. The F_1 score takes both false positives and false negatives into account and is the weighted average of precision and recall. F_1 is needed when we are seeking a

Table 3. Comparative Evaluation of proposed framework with its counter network

Model	precision	recall	f1-score
Proposed MMVT	**0.9986**	**0.9989**	**0.9987**
DenseNet-Freeze	0.945	0.955	0.948
ResNet-Freeze	0.972	0.972	0.976
Inception-Freeze	0.977	0.971	0.974
MobileNet-Freeze	0.953	0.95	0.95
DenseNet-Fine-tuned	0.887	0.883	0.883
ResNet-Fine-tuned	0.906	0.892	0.894
Inception-Fine-tuned	0.903	0.907	0.905
MobileNet-Fine-tuned	0.887	0.872	0.876
Khan etl.al. [7]	0.810	0.820	0.830
Qayyum et al. [15]	0.994	0.993	0.992

Table 4. Comparative Analysis (Average) of proposed MMVT framework with its counter network (ResNet, SqueezNet, Inception, MobileNet, DenseNet, Deep-CNN, and ShallowNet

	Avg. Accuracy	Avg Precision	Avg Recall	Avg F1score
Proposed MMVT	**99.72**	**99.85**	**99.82**	**99.72**
SqueezNet	86.92	89.25	82.20	84.46
DensNet121	87.65	86.95	86.94	86.78
ResNet101	90.37	89.98	84.28	86.41
Inception_V3	89.27	90.26	88.83	89.45
MobileNet-v2	87.32	85.40	88.68	86.71
Deep-CNN [7]	77.23	69.43	77.88	81.44
Qayyum et al. [15]	99.06	99.16	99.22	99.19

Table 5. Comparative Analysis (Average) of proposed MMVT with its counter networks (DenseNet, ResNet, MobileNet, and SqueezNet using Freeze based fine tuning

	Accuracy	precision	Recall	F1score
Proposed MMVT	**99.72**	**99.85**	**99.82**	**99.72**
SqueezNet	86.92	89.25	82.20	84.46
DensNet121	87.65	86.95	86.94	86.78
ResNet101	90.37	89.98	84.28	86.41
Inception_V3	89.27	90.26	88.83	89.45
MobileNet-v2	87.32	85.40	88.68	86.71
Deep-CNN [7]	77.23	69.43	77.88	81.44
ShallowNet Qayyum et al. [15]	99.06	99.16	99.22	99.19

balance between precision and recall. It is calculated as $F_1 = 2\frac{R \times P}{R+P}$. Specificity is the proportion of actual true negatives that were correctly predicted by the model that can be computed as and $Specificity = TN/TN + FP$ respectively.

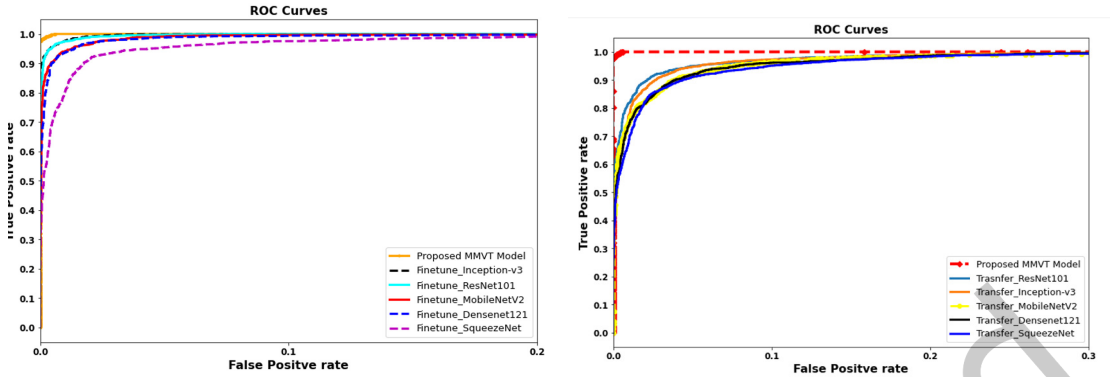


Fig. 7. Diagnostic performance Recall (left)- fine-tuning and (right)-Transfer learning

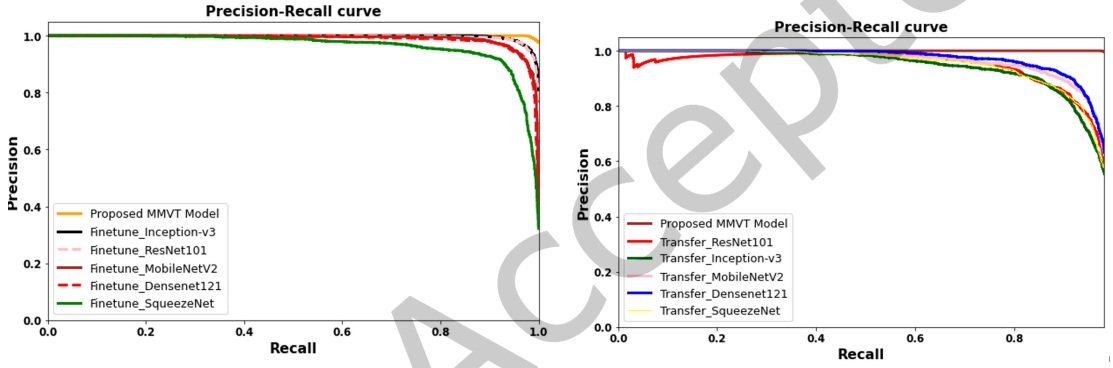


Fig. 8. Diagnostic performance Precision (left)- fine-tuning and (right)-Transfer learning

4.4 Results and Discussion

Design of a user Interface majorly focus on the goals that are specific to an application. Even though years of research on emotions, there is still a little understanding of emotions that how our interaction impacted by our emotions. In the recent years, emotions identification has shown significant development, however, most of this work focus either adult or posed expression whereas emotion recognition in children facial expression have received less attention considerably. Besides, pose-invariant occlusion-robust expression recognition in children in real-world scenarios have received comparatively less attention. Understanding emotion recognition deficit at early age may impact the our life and can result in poor social functioning. In this section, we present the experimental detail, evaluation of proposed MMVT and compartive analysis with state of the art methods on the spontaneous expression dataset. Figure 7, Figure 8, Table 3, Table 4, and Table 5,illustrate the experimental analysis of MMVT and its counter network. In order to generalize the gain in performance of proposed MMVT, we have performed cross-validation (10-fold) and compared the results using different evaluation measures.

Table 3 describes the performances of our approach for different strategical choices. We have trained different transformers in parallel to extract contextualized representations which are then forwarded to fully connected layer. First, we have used same network (twice) parallely to learn conextualized representation from the last

layers of transformers, however, it showed poor performance. To improve the performance, we have used two different models on RGB (vit_base_patch16_224, vit_base_patch16_224_miil) on RGB and Hematoxylin and eosin video in parallel to learn different features from each model. Finally, we have performed pairwise concatenation the learnt features form multiple transformer and passed to the fully connected layer at the end to concatenate the features of each individual transformers flowed by classification.

We have conducted extensive experiments on a benchmark child facial expression dataset (spontaneous child computer interaction) that showed significantly better performance in comparison to the state-of-the-art methods of CNN-based deep learning models. The CNN-based pre-trained models have been used as transfer learning and finetuning models on the child facial expression dataset. Various performance metrics have been used to validate the performance of proposed and existing deep learning models. To visualize the extracted contextualized representation, we have used t-distributed stochastic neighbor embedding visualization to visualize the discriminative feature in lower dimension space and probability density estimation (T-SNE-based feature dimension reduction), which is helpful to visualize the discriminative feature in lower dimension space for all child facial expression dataset classes. The probability density estimation showed the prediction capability of our proposed model. Figure 9 visualize the discriminative feature in lower dimension space that validate the robustness of the proposed framework in learning the contextualized representations. The probability density plot is based on the ground truth and predictions of proposed MMVT and pre-trained models as shown in figure 10. We can observe that the proposed MMVT framework has better discrimination power than its counter network. Besides, we further provided a probability density plot is based on the ground truth and predictions of proposed MMVT and pre-trained models.

Figure 10 and Figure 11 illustrate the color pattern visualization of the proposed model could better understand the weights activation used for model overfitting interpretation and could be helpful for clinical applications. The attention regions correspond to the right features shown in the activation map for a few samples. We have discussed two cases discussed in Figure 10 and Figure 11 to visualize the attention activation map. We can observe that the proposed MMVT provided significantly better attention to the facial expression areas for most of the images. It would be better to see the feature maps activation of the respective class where the features provide more attention for the good features. The red and blue color shows the most important features in a different region of the predicted model.

5 CONCLUSION

Unlike, most of the existing work that focus majorly facial expression recognition is based posed expression (fake or disguised inner feeling), we focused on spontaneous facial expressions that captured in an unconstrained environment. We presented multimodal RGB video and Hematoxylin and eosin video-based transformer-based framework to extract contextualized representations from RGB video sequence and Hematoxylin and eosin video sequence and then use these representations to predict user's emotions. We have performed a late fusion through pairwise concatenations of contextualized representations using the cross-feature fusion technique, which results in robust and efficient contextualized representations. We have used different pre-trained models with different network structures to improve the performance. Experiments were conducted on benchmarks child spontaneous dataset that showed significant improvement in performance as compared to benchmark spontaneous facial expression recognition. One of the major challenges is that emotions in human-computer interactions are rarely discrete and rarely limited to the six basic emotions. However, most of the dataset covers only six basic emotions. In the future, we will work on the development of datasets that consist of more emotions to address the limitation of current datasets.

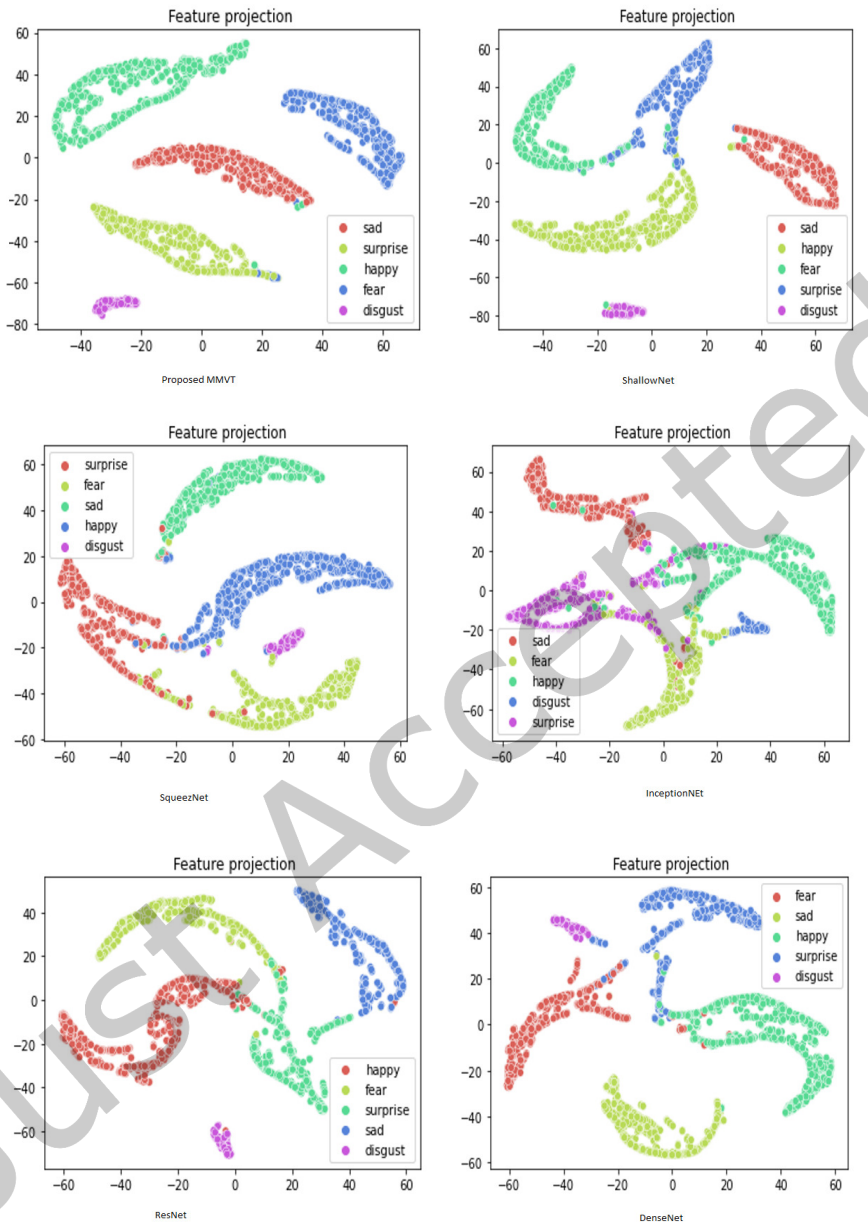


Fig. 9. Diagnostic performance (a) Precision Recall (b) ROC plot on EEG dataset

REFERENCES

[1] Amani Albraikan, Basim Hafidh, and Abdulmotaleb El Saddik. 2018. iAware: A real-time emotional biofeedback system based on physiological signals. *IEEE Access* 6 (2018), 78780–78789.

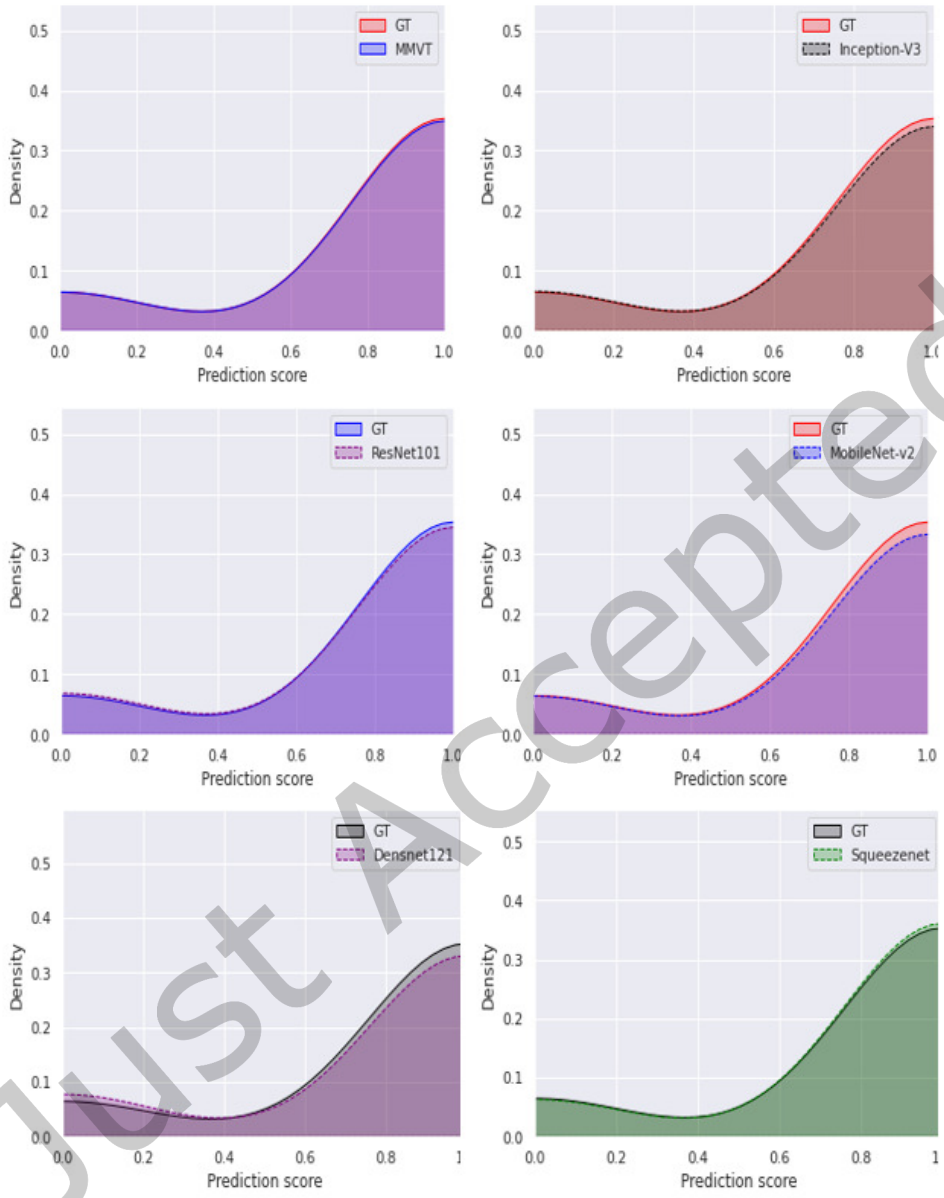


Fig. 10. The probability density plot is based on the ground truth and predictions of proposed MMVT and pre-trained models.

[2] Amani Albraikan, Diana P Tobón, and Abdulmotaleb El Saddik. 2018. Hyper-Parameter Optimization for Emotion Detection using Physiological Signals. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 836–841.

[3] Amani Albraikan, Diana P Tobón, and Abdulmotaleb El Saddik. 2018. Toward user-independent emotion recognition using physiological signals. *IEEE sensors Journal* 19, 19 (2018), 8402–8412.

Fig. 11. Attention activation map for child facial expression detection.

Fig. 12. Attention activation map for child facial expression detection.

[4] Russell Beale and Christian Peter. 2008. The role of affect and emotion in HCI. In *Affect and emotion in human-computer interaction*. Springer, 1–11.
[5] Girum G Demisse, Djamila Aouada, and Björn Ottersten. 2018. Deformation-based 3d facial expression representation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1s (2018), 1–22.
[6] Huang Gu, Qiong Chen, Xiaoli Xing, Junfeng Zhao, and Xiaoming Li. 2019. Facial emotion recognition in deaf children: evidence from event-related potentials and event-related spectral perturbation analysis. *Neuroscience letters* 703 (2019), 198–204.
[7] Rizwan Ahmed Khan, Arthur Crenn, Alexandre Meyer, and Saida Bouakaz. 2019. A novel database of children's spontaneous facial expressions (LIRIS-CSE). *Image and Vision Computing* 83 (2019), 61–69.
[8] Yelin Kim and Emily Mower Provost. 2015. Emotion recognition during speech using dynamics of multiple regions of the face.
[9] Natalia Kucirkova, Cecilie Evertsen-Stanghelle, Ingunn Studsrød, Ida Bruheim Jensen, and Ingunn Størksen. 2020. Lessons for child–computer interaction studies following the research challenges during the Covid-19 pandemic. *International journal of child-computer interaction* 26 (2020), 100203.
[10] Shiguang Liu, Huixin Wang, and Min Pei. 2022. Facial-expression-aware Emotional Color Transfer Based on Convolutional Neural Network. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 1 (2022), 1–19.

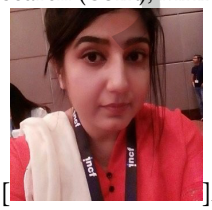
[11] Juan Manuel Mayor-Torres, Sara Medina-DeVilliers, Tessa Clarkson, Matthew D Lerner, and Giuseppe Riccardi. 2021. Evaluation of Interpretability for Deep Learning algorithms in EEG Emotion Recognition: A case study in Autism. *arXiv preprint arXiv:2111.13208* (2021).
[12] Vida Mehdizadehfar, Farnaz Ghassemi, Ali Fallah, and Hamidreza Pouretemad. 2020. EEG study of facial emotion recognition in the fathers of autistic children. *Biomedical Signal Processing and Control* 56 (2020), 101721.
[13] Raja Majid Mehmood and Hyo Jong Lee. 2016. A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns. *Computers & Electrical Engineering* 53 (2016), 444–457.
[14] Yu Miao, Haiwei Dong, Jihad Mohamad Al Jaam, and Abdulmotaleb El Saddik. 2019. A deep learning system for recognizing facial expression in real-time. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 2 (2019), 1–20.
[15] Abdul Qayyum, Imran Razzak, Nour Moustafa, and Mona Mazhar. 2022. Progressive ShallowNet for Large Scale Dynamic and Spontaneous Facial Behaviour Analysis in Children. *Image and Vision Computing* (2022), 104375.
[16] Caifeng Shan, Shaogang Gong, and Peter W McOwan. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing* 27, 6 (2009), 803–816.
[17] S Voeffray. 2011. Emotion-sensitive human-computer interaction (HCI): State of the art-Seminar paper. *Emotion Recognition* (2011), 1–4.
[18] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. 2020. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing* 29 (2020), 4057–4069.
[19] Xueping Wang, Yunhong Wang, and Weixin Li. 2019. U-Net conditional GANs for photo-realistic and identity-preserving facial expression synthesis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–23.
[20] Huei-Fang Yang, Bo-Yao Lin, Kuang-Yu Chang, and Chu-Song Chen. 2018. Joint estimation of age and expression by combining scattering and convolutional networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1 (2018), 1–18.
[21] Anbang Yao, Dongqi Cai, Ping Hu, Shandong Wang, Liang Sha, and Yurong Chen. 2016. HoloNet: towards robust emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 472–478.
[22] Guanghao Yin, Shouqian Sun, Dian Yu, Dejian Li, and Kejun Zhang. 2022. A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 3 (2022), 1–23.
[23] Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*. 222–237.
[24] Wei Zhang, Ting Yao, Shiai Zhu, and Abdulmotaleb El Saddik. 2019. Deep learning–based multimedia analytics: a review. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 1–26.
[25] Zhaoxin Zhang, Changyong Guo, Fanzhi Meng, Taizhong Xu, and Junkai Huang. 2020. CovLets: A Second-Order Descriptor for Modeling Multiple Features. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 1s (2020), 1–14.
[26] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 915–928.
[27] Ruicong Zhi, Markus Flierl, Qiuqi Ruan, and W Bastiaan Kleijn. 2010. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 41, 1 (2010), 38–52.
[28] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. 2012. Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2562–2569.

[ ]Abdul Qayyum Dr. Qayyum did his PhD in Electrical and Electronics Engineering from Universiti Teknologi PETRONAS, Malaysia 2018. He developed deep learning based algorithms for depth estimation of vegetation, trees near power lines for Tenaga Nasional Berhad (TNB) and Sabah Electric Supply Berhad (SESB) under the ministry of Green , Water and Technology (KeTTHA) Malaysia. He also developed a prototype for Vital signs (heart rate, breathing rate, Sp02) estimation and assessment of stroke and Arterial fibrillation (AF) using face video analytic based on deep leaning models. Earlier, he had completed his Bachelors in Computer Engineering and Master in Electronic Engineering from Pakistan. Besides, he gained one year industrial experience while working as a BSS engineer for Huawei, Pakistan. He had also taught several courses under electrical, specifically, signal processing domain for 7 years in various public and private universities in Pakistan. He was working

as a research scientist in CISIR, UTP for less than one year and was developed deep learning algorithms for brain signal (EEG) classification and reconstruction, remote sensing image segmentation and biomedical image analysis. Currently, he is associated with burgundy university France as a Post Doc researcher and working on the cardiac MRI images using deep learning approach. He is also working as a consultant in various projects involving deep learning models in Big Data, Vital Sign Estimation, IoT and BCI applications

]Imran Razzak Imran Razzak is a Senior Lecturer in Human-Centered AI and Machine Learning at the School of Computer Science and Engineering, University of New South Wales, Sydney. He is an associate editors/guest editor of several journals such as IEEE TCSS, IEEE JBHI, IEEE TII, etc. His area of research includes machine learning and NLP with its application to a broad range of topics, particularly deep learning, big data analytics, healthcare, and cyber security. His research mainly focuses on the healthcare sector, and he is passionate about making the healthcare industry a better place through emerging technologies.

]M. Tanveer is Associate Professor and Ramanujan Fellow at the Department of Mathematics of the Indian Institute of Technology Indore. Prior to that, he worked as a Postdoctoral Research Fellow at the Rolls-Royce@NTU Corporate Lab of the Nanyang Technological University, Singapore. His research interests include support vector machines, optimization, machine learning, deep learning, applications to Alzheimer's disease and dementia. He has published over 80 referred journal papers of international repute. His publications have over 2150 citations with h index 26 (Google Scholar, April 2021). Recently, he has been listed in the world's top 2% scientists in the study carried out by Stanford University, USA. He is currently the Associate Editor - IEEE Transactions on Neural Networks and Learning Systems (Feb. 2022 -), Associate Editor - Pattern Recognition, Elsevier (Nov 2021 -), Action Editor - Neural Networks, Elsevier (Jan 2022 -), Board of Editors - Engineering Applications of Artificial Intelligence, Elsevier (Jan 2022 -), Associate Editor - Neurocomputing, Elsevier (Jan 2022 -), Associate Editor - Cognitive Computation, Springer (Jan. 2022 -), Editorial Board - Applied Soft Computing, Elsevier (Jan 2022 -), International Journal of Machine Learning and Cybernetics, Springer (July 2021 -). He has also co-edited one book in Springer on machine intelligence and signal analysis. Tanveer is currently the Principal Investigator (PI) or Co-PI of 11 major research projects funded by Government of India including Department of Science and Technology (DST), Science Engineering ResearchBoard (SERB) and Council of Scientific Industrial Research (CSIR), MHRD-SPARC, ICMR.



]Moona Mazher Moona Mazher received the bachelor's degree in telecommunication engineering from Pakistan in 2010 and the master's degree in electrical and electronics engineering with specialization in biomedical sciences (neuroscience) from Universiti Teknologi PETRONAS, Malaysia, in 2017. She had published

numerous journal and conference papers during the master's degree. Her research interests include biomedical signal processing, machine learning, and the optimization of deep learning algorithms.