2023 International Conference on Innovations in Engineering and Technology (ICIET) | 979-8-3503-0999-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIET57285.2023.10220883

A Study on Human Emotion Recognition Techniques

Kavitha K V
Department of Computer Science and Engineering
Annamalai University
Chidambaram
kaviabhilash123@gmail.com

Sudha L R
Department of Computer Science and Engineering
Annamalai University
Chidambaram
sudhaselvin@gmail.com

Jayasudha J S
Department of Computer Science
Central University of Kerala
Kasaragod
jayasudhajs@gmail.com

***Abstract*— Emotions are essential for humans and play an important role in cognition. Emotion is frequently linked to logical decision-making, outlook, interactional behavior, and, to a lesser extent, human intelligence. With a growing interest in ensuring "emotional" interactions amid machines and humans, there is a need for reliable solutions in recognizing human emotional states. Humans typically express their emotions through a variety of non-linguistic and indirect methods. Emotion detection is accomplished primarily through the use of six distinct strategies: facial emotion recognition, physiological signal recognition, speech variation, emotion recognition from audio and videos, and emotion recognition from EEG signals on datasets like RAVDESS, JAFFE, CK+, TESS, DEAP, etc. In general, these methods identify seven basic emotions. In addition, we compared different ways to recognize emotions in humans. Deep convolutional neural networks produced the best results for Facial Emotion Recognition, physiological signals, audio, video, and EEG, emotion recognition through speech with accuracies of 97.83%, 85%, and 70%, 75%, 80% respectively.**

***Keywords*— Human Emotion Recognition, Facial Emotions, Speech, Physiological Signals, Electroencephalogram Signals (EEG), Audio, Video.**

I. INTRODUCTION

Emotions are necessary for logical decision-making, perspective, knowledge acquisition, and a variety of other functions. Granting devices, the ability to understand human emotions would thus improve and deepen Human Computer Interaction (HCI). It is an interdisciplinary field that includes, to name a few, robotics, emotion recognition, data mining, and HCI. Affective computing is a science that aims to replicate, as well as process, identify, and comprehend emotional states. For instance, if the computer is aware of the student's emotional state and offers the appropriate learning, the student's receptivity during online learning will be significantly increased. When a psychologist is aware of the patient's emotional state, they can quickly diagnose the illness.

Speech is a complex signal that contains information more about speaker, the comment, the vocabulary, the emotion, and much more. The vast majority of current speech systems perform well when processing studio-recorded, neutral speech, but badly when accessing emotional speech. This happens because it is difficult to model and categorise the emotions expressed in speech. When emotions are present, speech becomes more natural. From a computer's point of view, classifying or separating different emotions can be viewed as understanding speech emotions.

Researchers working on brain computer interface (BCI) devices have recently become interested in detecting emotional changes in EEG signals. However, compared to EEG signals with longer durations for emotion recognition, the majority of earlier works did not analyze short-duration EEG signals. The key problems in this area of research are trying to identify non-linearity in EEG signals, as well as selecting the most efficient emotion-inducing stimuli. It is possible to create neural networks that recognize emotions in audio. The external presentation, or facial expression, is perceptible but easily corrected by the environment. EEG is a highly anticipated tool for emotion recognition due to its widespread use as a Physiological Signal (PS) closely linked to emotions and effectively managed to capture from brain area expression. Skin temperature, EEG, EMG, blood volume, ECG, EOG, and other PS are examples. Some research findings used physical data to conduct emotion recognition due to the stability of the PS. As it is assigned from the brain cortex, EEG is regarded as promising among physical signals. Facial recognition technology is now widely used in a variety of applications, including security-related apps, digital services, educational facilities, and to name a few. Fig. 1 depicts the domain categorization of various Emotion Recognition Techniques. The objective of this study is to conduct a thorough examination of significant audio, speech, facial, physiological, and video, EEG based HER methods that have been implemented over previous decade.

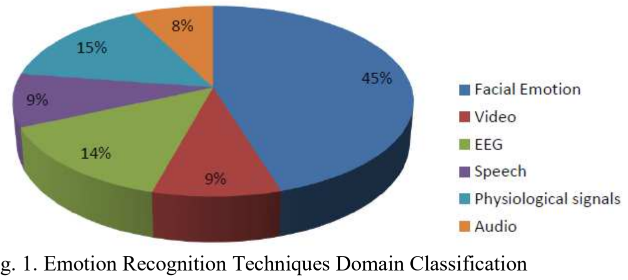

Fig. 1. Emotion Recognition Techniques Domain Classification

Face emotion recognition is primarily accomplished through the use of two methods: feature-based techniques and model-based techniques like Gabor wavelets, Weber Local Descriptors (WLD), facial landmarks, Active Units (AUs). The paper is arranged as follows- second section is the methodology and third section is the related work which discusses the different techniques used in ER through facial are discussed in fourth section with comparison for different ER domains. The last section states the conclusions and scope of future work.

II. METHODOLOGY

ER methods are implemented to identify emotions from speech, audio, video, face images, EEG, and PS. A CNN method is used to identify and classify emotions based on input signals. The signals come from various databases and are classified into following categories: normal, excitement, anxiety, unhappiness, frustration, amazement, and disgust. The strategy begins with gathering datasets from open source online, followed by pre-processing the data to prepare it for data processing.

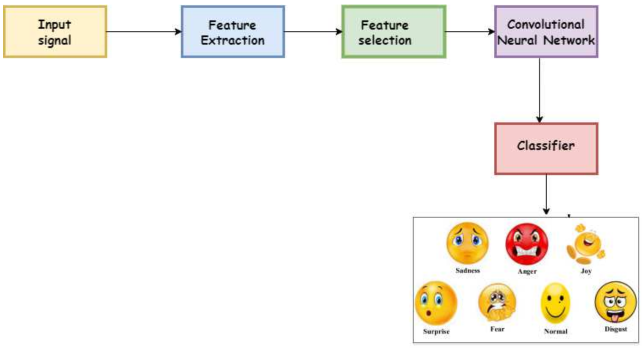

Fig. 2. Block Diagram of Human Emotion Recognition

Fig. 2 shows the block diagram of ER techniques. The feature extraction method has the ability to create more generally applicable models, and reduce computational complexity. The feature selection method selects relevant attributes from a set. The mean operation selects its most relevant attributes from a set of combined features. CNN's emotion recognition network is fed the extracted features. CNN's feature extraction process begins with the convolutional layer. This layer procedure combines two different input sets: a data matrix and filtration. As input and output, it has an interconnected layer. This study's CNN receives the signals from various datasets.

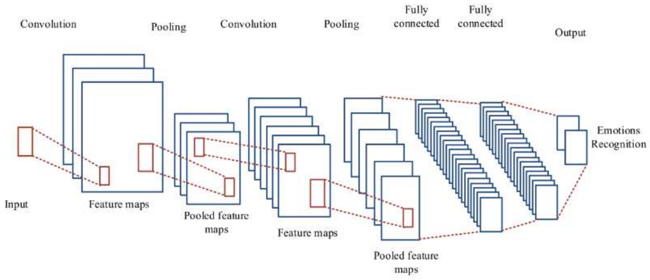

Fig. 3: Architecture of CNN

Fig. 3 depicts the architecture of the CNN classifier. The modeled CNN's input is made up of neurons that collect highly standardized fragment sizes. The built CNN model comprises three key operations: activation, pooling, and kernels. Furthermore, the established model is made up of fixed-size kernels. The kernel would then be pre-trained using an unsupervised learning approach based on patches. The feature map is taken into account by the convolutional layer. The most common model for feature pooling is down-sampling, which is conducted using the mean operation.

III. RELATED WORKS

A. Emotion Detection Based on Facial Images

Dynamic FER multi-attention modules were used for emotion detection. The data sets CK+ and eNTERFACE05 were used in this case. This method produced three distinct attention functionalities: spatial, channel, and frame. Highest accuracy rates for the datasets are 89.52% and 88.33%, respectively. Byun et al. [1] proposed a weighted integration method. The RAVDESS dataset was used to recognize emotions from a facial image pattern along with CNN and LSTM. This model significantly improved accuracy by 87.11%. Deep Learning (DL) approaches for FER were proposed to improve accuracy. The experiment was carried out using CK+ database, and the average accuracy was 96%. This study addresses ER using transfer learning approaches. Facial Image Threshing (FIT) was introduced for FER with 86.95% accuracy. Arora et al. [2] investigated an automatic system that recognizes various emotions connected on the face. The JAFFE dataset was employed in this case and the proposed technique achieves a maximum classification rate of 94.97%.

Mehendale *et al.* [3] proposed a novel FER technique based on CNN. This FERC model has 96% accuracy rate. Verma et al [4] proposed a hybrid DL model for FER. The suggested model was trained, and it has an accuracy of 97.07% in the FER2013 dataset and 94.12% in the JAFFE dataset. A framework made up of two ML algorithms that are used for detection and classification. AdaBoost cascade classifiers were used in this case. This method provided accuracy of 57.7% in SFEW dataset and 59.0% accuracy in RAF dataset. Lekshmi et al [5] developed FER feature descriptor based on Histogram of Oriented Gradients (HOG) and LBP for determining facial characteristics from CK+ and JAFFE datasets. The recognition accuracy of the proposed work is 97.66%. Saurav et al. [6] evaluated DCNN models' performance on various datasets and model recognition accuracy is 87.16%. Face-Sensitive CNN (FS-CNN) for HER was provided by Said et al [7] to improve performance. FS-CNN is made up of two processes: patch cropping and CNN. FS-CNN provided 94.9% accuracy. HER model based-on meta learning across pose, occlusion and illumination provided 90% accuracy.

In face feature extraction for ER, multilevel stationary bi orthogonal wavelet transform provided 56.5% accuracy. Multi-branch deep radial basis function networks performed FER with 99.64% accuracy. Fusion of key points descriptor and texture features were introduced for ER using facial expression. This method involves three phases of ROI extraction, duplex feature fusion, and classification. The suggested work enhances the recognition rate by about 97%, 88%, 86%, and 93%, and decreases the misclassification rate by approximately 1.4%, 7.6%, 6.6%, and 2.7%. Evolutionary algorithms were used for Image-based ER. The overall accuracy is obtained as 98.67%. A GA-based Linear Discriminant Analysis (LDA) classifier and a PCA-based feature selection technique are employed. The suggested method outperforms existing genetic-based feature selection algorithms as well as linear-based dimensionality reduction algorithms. Haar Cascade method and NN approach were used to detect the input image, a face, and a mouth in order to recognize all types of emotions with an accuracy of 82.58% in JAFFE dataset. Hybrid DNN model has an accuracy of 92.07% in the MMI dataset and of 94.91% in the JAFFE dataset.

For audio-visual emotion recognition, multi regression and the Ridgelet transform were used. This study made use of the eNTERFACE database. For the speech and face modalities, two different extreme learning machine classifiers are used. For bimodal input data, speech, and face, the proposed technique achieves accuracies of 85.06%, 64.04%, and 58.38%, respectively. CNN was implemented on CK+ dataset provided accuracy of 95.76%, while ADFES dataset had an accuracy of 88.73%. In the preceding sections, we evaluated and contrasted all of the methods and approaches in the same field, as well as the methodological approaches of the various ER methods. We compared model-based and feature-based methods for recognizing facial emotions. Fig. 4 illustrates the outcomes of FER approaches.

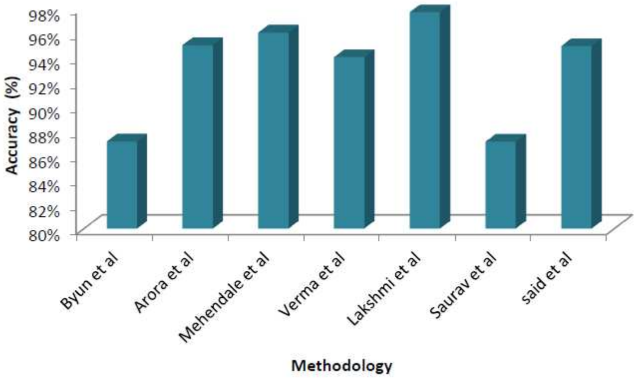

Fig. 4. Accuracy chart for ER in facial images

B. Emotion Recognition Based on EEG

EEG-oriented DL approach was developed for ER. The binary gray wolf optimizer was used to carry out the feature selection task. This model for ER was developed using the DEAP dataset with average accuracy is 96.87%. Asymmetric Map (AsMap) was developed for emotion classification. The CNN model's AsMap has the highest accuracy rate of 97.10%. DEAP and SEED Dataset are the two most commonly used EEG emotion datasets. Suhaimi et al [8] proposed virtual reality for ER. This model has a classification accuracy of 85.01%. Joshi et al [9] proposed BiLSTM classifier for emotional state classification. This model provided accuracy of 74% in the DEAP database and 45% in the SEED database. CNN classifier was implemented in GAMEEMO dataset that provided accuracy rate of 82.32%. Islam et al [10] investigated the CNN model's ability to recognize emotion from EEG signals. In terms of accuracy, they achieved 78.22% for valence and 74.92% for arousal. Liu et al [11] suggested a feature fusion that provided accuracy of 81%. ScalingNet was proposed by Hu et al [12]. On DEAP and AMIGOS datasets, accuracy obtained are 71% and 73.89% respectively.

Nonlinear Entropy Metrics was developed for ER. SVM classifier is used, and the accuracy of Quadratic Sample Entropy (QSE) and Permutation Min-Entropy (PME) is 96.39%. Multivariate empirical mode decomposition was presented for ER. The DEAP data set's accuracy is 77%. Based on EEG signals, an automated model for identifying emotions and CNN provide precision of 94.09%. Hybrid intersections were implemented to identify emotions using EEG. The classification accuracy of MSVM is 89.76%.

Gao et al [13] suggested a procedure for identifying fusion features in EEG-based ER. A concatenated feature extraction method is employed to classify emotions. Experimental results based on SVM and RVM classifiers have average accuracy of 89.17% and 91.18%, respectively. Learning CNN characteristics from DE features. For SEED dataset, this model provided average accuracy of 90.41%. Mert et al [14] proposed Multivariate Empirical Mode Decomposition (MEMD) with 75% accuracy.

Using EEG signals, the domains of ER are evaluated. These signals necessitate extensive pre-processing, making feature extraction challenging. As a result, these are the methods for recognizing emotions that are used the least. Gao et al have the highest accuracy, as shown in Fig. 5, while the majority of the remaining methods have accuracies ranging from 69 to 90%. EEG signals are PS that record brain activity. Before feature extraction and selection, EEG signals, like ECG signals, require extensive pre-processing. The accuracy of the graph for ER via EEG signals with SVM classifiers is the lowest.

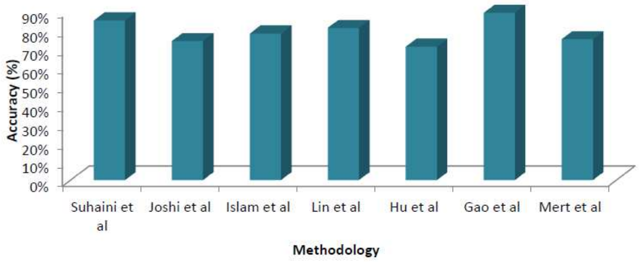

Fig. 5. Accuracy chart for EEG based ER

C. Emotion Recognition Based on Speech

For extracting Speech ER (SER) features, two-way feature extraction and deep transfer method were used. RAVDESS dataset has accuracy of 81.94%, while TESS dataset has accuracy of 97.15%. Deep CNN for SER was developed and the accuracy was ascertained to be 92.76%. BLSTM Directional Self-Attention (BLSTM-DSA) was combined with Cepstrum coefficients for ER. The model's accuracy is 93.8931%. Cat Swarm Optimization (CSO) algorithm was introduced for recognizing speech emotions. SVNN categorized the emotions that are provided. ECSO-SVNN has an accuracy of 96%.

Yildirim et al [15] implemented a technique for selecting features for metaheuristic algorithms. The feature selection methods are evaluated using the USC IEMOCAP and EMO-DB databases. The highest accuracy obtained using the SVM classifier is 87.66%. Singh et al. [16] suggested a multimodal hierarchical approach for extracting SER from audio and text. Spectral, Prosody, and VQ features are the three major categories of features for emotion classification tasks. On audio features, the accuracy of a hierarchical DNN-based classifier was 81.2%. The proposed data model provided 81.7% accuracy. ML Algorithms for SER were proposed by Prasanth et al [17]. This model will employ two feature approaches: Prosodic (long-term) features and spectral features. The accuracy of the SER classifier is 72%. Abdul et al [18] achieved near-perfect classification accuracy of 96.3% using statistical approach. AM-FM model was developed for speech analysis. SVM-based classifier provided recognition rate of 91.16%. Ren et al [19] proposed an original multi-modal correlated network and extracted video and audio features using 3D-CNN and 2D-CNN respectively. The model accuracy is calculated to be 60.59 for the AFEW Dataset.

Multi-level decision concept was combined with the DL model for SER. The concept of multi-level categorization is mostly realized through the use of tree structure. For multi-level SER, Zheng et al [20] proposed an ensemble model. The IEMOCAP dataset is used in this case. Weighted Accuracy (WA) and Unweighted Accuracy (UA) were used as indicators to avoid the impact of various emotional imbalances with 75% accuracy. Segment Repetition based on High Amplitude (SRHA) was introduced for improving SER with an accuracy of 98.25%. Semi-Supervised Ladder Networks (SSLN) provided accuracy of 59.7%. Poorna et al [21] proposed a multistage classification scheme for enhancing SER with Emo-DB database. The average recognition accuracies with ELM and SVM are 88.2% and 85.5%, respectively. Stacked deep auto-encoder model with a wavelet kernel sparse classifier was developed for SER. The softmax classifier has a 14.7% recognition rate. For SER, altered brain emotional learning model was introduced. The ANFIS architecture, which contains both ANN and fuzzy logic was used. It is discovered that different models have the lowest accuracy for emotion recognition through speech signals, while method of Yildirim et al has the highest accuracy of 87.76% as shown in Fig. 6. As per chart, the accuracy of identifying emotion via speech signals has significantly improved over time.

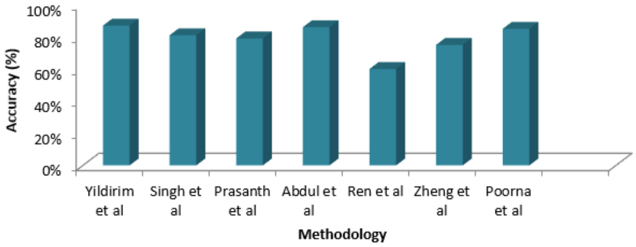

Fig. 6. Accuracy chart for Speech ER

D. Emotion Recognition Based on Physiological Signals

Jimenez et al [22] proposed ML model for ER from Physiological Signals (PS). This study proposed a model for categorizing enjoyment, unhappiness, and neutral emotions. The DEAP dataset was used to test the methodology. The accuracy of SVML was determined to be 92%. Ali et al [23] proposed a novel specific topic HER system for emotion recognition from PS. On the MAHNOB dataset, CNN model performs significantly better, with 89.38% accuracy. Nasoz et al [24] described a novel approach to improving ER presence technologies from PS. They present a multimodal affective user interface prototype. ER results using the KNN algorithm with 70% accuracy. Ayata et al [25] suggested a new ER algorithm based on multimodal PS using DEAP dataset. Emotions are represented using dimensional models and category models. A light source is used in Photo Plethysmography (PPG), and the amount of transmitted or reflected light is measured. The proposed model's accuracy was determined to be 73.08 %.

Ren et al [26] proposed an Asymmetry Index (AsI) and an echo state network based on PS for ER. The DEAP provided 78.2% accuracy. Zhang et al [27] used Augsburg Biosignal Toolbox (AuBT) to evaluate PS related to emotion perception. This model has classification accuracy of 93.42%. Parallel stacked autoencoders were used for ER from PS using DEAP database. The proposed model's average accuracy was calculated to be 93.6%. Fang et al [28] postulated hierarchical fusion of visual and PS recognition. They proposed a sequential fusion of CNN and neural aggregation network. In this case, the MAHNOB-HCI dataset is used with 69.21% accuracy. Collaborative learning for ER from peripheral PS was developed. In the DEAP dataset, the proposed method achieves 98.6% accuracy. Fig. 7 depicts the performance of PS recognition. The classification accuracy and findings achieved using unified psycho PS by various methods are shown. The results show that only using psycho PS to recognize some emotions actually improves, but the performance is still satisfactory.

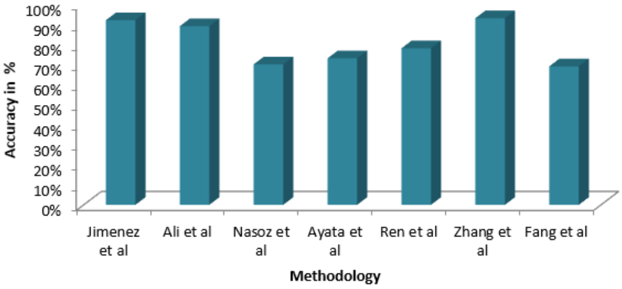

Fig. 7. Accuracy chart for Physiological signals ER

E. Emotion Recognition Based on Audio

Cunningham et al [29] proposed supervised ML for Audio Emotion Recognition (AER). The ANN models explained 64.4% and 65.4 % of the variation in the prediction of arousal and valence, respectively. The IADS dataset provided categorization rate of 43.7% for arousal. Deep audio embeddings for AER were proposed by Koh et al [30]. They created a number of multi-class classifiers that predict emotion meanings in music using deep audio embeddings. The proposed model's accuracy is found to be 88%. Patel et al. [31] proposed conventional AER auto encoders with accuracy 91.92%. Middya et al [32] proposed multimodal ER by fusing audio-visual modalities at the model level. HER from audio signals was proposed by Chennoor et al [33] and obtained of accuracy of 77.33%. For audio compression, Reddy et al [34] proposed multi algorithm fusion. The Automatic Emotion Recognition System (AERS) is a technique for tracking and diagnosing emotional/psychological conditions within a unit. Classification of various emotional states is used to examine the retrieved features. The accuracy of this model is 94% by employing a multi-resolution strategy and estimating frequency as well as time data. Fig. 8 depicts a comparison between ER techniques based on audio.

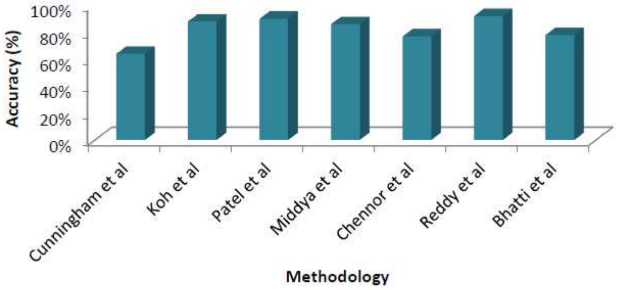

Fig. 8: Accuracy chart for Audio ER

Multimodal fusion with DNN was developed for Audio-Video ER. A new DNN architecture was offered for estimating a subject's emotional state. Bhatti et al [35] suggested used brain signals to create audio music for HER. Music is thought to be an effective tool for arousing human emotions in this context. This model's accuracy is calculated to be 78.11%. SVM and K-NN classifiers are used. For audio-visual ER, Fuzzy Neural Network (FNN) was developed. SAVEE database's emotion recognition accuracy was 98.25%. The audio clips were extracted from the training dataset.

F. Emotion Recognition Based on Video

Chen et al [36] described a traditional KNN model which produced best classification accuracy of 39.70%. Wei et al [37] created a key frame extraction method for Video Emotion Recognition (VER). Ekman-6 and VideoEmotion-8 and user-generated video datasets have accuracy of 59.51% and 52.85%, respectively. HOMER, a cloud-based technology, was employed for video-based ER. F1-score of HOMER demonstrates 38% from the baseline. The dataset Video Titles in the Wild (VTW) is used in this method. Through the use of two smart phone applications, this method demonstrated the mobility and scalability of HOMER. DL-based classifier was proposed by Pandeya et al [38]. This model consists of a slow-fast network that employs filter convolution. This multimodal architecture achieved an accuracy of 74%.

Wang et al [39] proposed multimodal DL for ER using multiple psycho-PS and video. The Bio-Vid Emo database was used in this approach. The proposed model's SVM classification accuracy is 80.89%. Singh et al [40] proposed DNN for audio and video ER. For IEMOCAP dataset, the model's accuracy is calculated to be 71.75%. Dresvyanskiy et al. [41] proposed transfer learning framework for ER. They used two different weighted score fusion techniques. The test set's performance metric yields 42.10%. Deep networks based on 1D CNN and LSTM are utilized for transfer learning. Multimodal Fusion with DNN was introduced for ER. A DNN architecture is created to estimate a subject's emotional state. Zhalehpour et al [42] proposed three new peak frame selection approaches for ER that take advantage of the video channel. This model has an audio-visual accuracy of 78.26%.

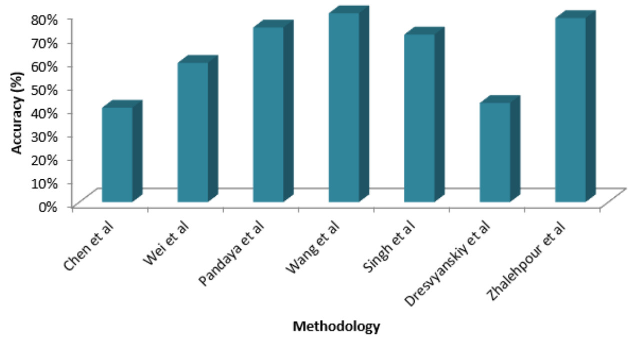

Fig. 9: Accuracy chart for Video ER

Fig. 9 compares the performance of different ER methods for videos on different datasets. Initially, the training set's videos are divided into frames at a 24 FPS rate. Now, an image labelled database of images is manually created from video clip frames. The trained model predicted the emotion from each frame. The emotion that repeats in the most frames is the overall emotion carried throughout the video clip. When compared to the baseline, the method used in Zhalehopour et al achieved 78% accuracy.

IV. DISCUSSIONS

In this study, a thorough examination has been conducted on the significant audio, speech, facial, physiological, and video, EEG based HER methods that have been implemented over previous decade. Majority of the literatures used facial image as the domain for HER. Several sophisticated algorithms were developed in the facial image-HER domain. Fig. 10 depicts the accuracy of HER using various signal modalities. From the literature studied, it has been found that FER algorithms provide the highest ER accuracy of 97%. Video based HER algorithms provide the lowest accuracy of 70%. It can be implicated that facial images are the right domain for HER. Novel algorithms need to be implemented in FER domain for the improvement of accuracy and other parameters.

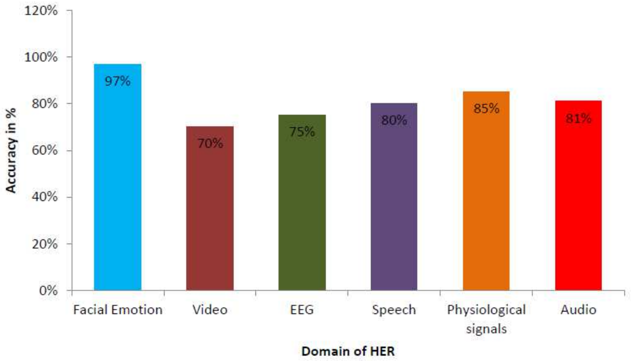

Fig. 10. Accuracy chart for various HER Techniques

V. CONCLUSION

Emotions are highly important in the social life of human being. This study provides an evaluation and compile significant HER techniques that have been implemented. We now have a variety of methods at our disposal, ranging from those utilize a single neural model to those integrates various features, models, and classifiers. Grief, amaze, disgust, joy, anxiety, and anger are the six basic emotions that humans express. To assess the effectiveness of these methods, a proper investigation can be conducted. There are still few precisely merged hybrid models available. It is possible to develop more effective hybrid and merged methods for assessing human emotions. To assess the effectiveness of these methods, a detailed search can be conducted. Based on the study it is clear that facial ER is more accurate compared to other modalities. Novel algorithms can be developed in ER domain to enhance recognition accuracy.

REFERENCES

[1] S. W. Byun, and S. P. Lee, “Human emotion recognition based on the weighted integration method using image sequences and acoustic features,” *Multimedia Tools and Applications*, vol. 80, pp. 35871-35885, 2021.

[2] M. Arora, and M. Kumar, “AutoFER: PCA and PSO based automatic facial emotion recognition,” *Multimedia Tools and Applications,* vol. 80, pp. 3039-3049, 2021.

[3] N. Mehendale, “Facial emotion recognition using convolutional neural networks,” *SN Applied Sciences,* vol. 2, pp. 1-8, 2020.

[4] G. Verma, and H. Verma, “Hybrid-deep learning model for emotion recognition using facial expressions,” *The Review of Socio network Strategies*, vol. 14, pp. 171-180, 2020.

[5] D. Lakshmi, and R. Ponnusamy, “Facial emotion recognition using modified HOG and LBP features with deep stacked

autoencoders," *Microprocessors and Microsystems*, vol. 82, pp. 1038-1046, 2021.

[6] S. Saurav, R. Saini, and S. Singh, "EmNet: a deep integrated convolutional neural network for facial emotion recognition in the wild," *Applied Intelligence*, vol. 51, pp. 5543-5570, 2021.

[7] Y. Said, and M. Barr, "Human emotion recognition based on facial expressions via deep learning on high-resolution images," *Multimedia Tools and Applications*, vol. 80, pp. 25241-25253, 2021.

[8] N. S. Suhaimi, J. Mountstephens, and J. Teo, "A Dataset for Emotion Recognition Using Virtual Reality and EEG," *Big Data and Cognitive Computing*, vol. 6, pp. 16-28, 2022).

[9] V. M. Joshi, and R. B. Ghongade, "EEG based emotion detection using fourth order spectral moment and deep learning," *Biomedical Signal Processing and Control*, vol. 68, pp. 1027-1039, 2021.

[10] M. R. Islam, M. M. Rahman, M. A. Moni, "EEG channel correlation-based model for emotion recognition," *Computers in Biology and Medicine,* vol. 136, pp. 1047-1059, 2021.

[11] Y. Liu, and G. Fu, "Emotion recognition by deeply learned multi-channel textual and EEG features," *Future Generation Computer Systems*, vol. 119, pp. 1-6, 2021.

[12] J. Hu, C. Wang, Q. Jia, and J. Feng, "ScalingNet: extracting features from raw EEG data for emotion recognition," *Neurocomputing*, vol. 463, pp. 177-184, 2021.

[13] Q. Gao, Q. Wang, C. H. Wang, and Y. Song, "EEG based emotion recognition using fusion feature extraction method," *Multimedia Tools and Applications,* vol.79, pp. 27057-27074, 2020.

[14] A. Mert, and A. Akan, "Emotion recognition from EEG signals by using multivariate empirical mode decomposition," *Pattern Analysis and Applications*, vol. 21, pp. 81-89, 2018.

[15] S. Yildirim, Y. Kaya, and F. Kılıc, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Applied Acoustics*, vol. 173, pp. 1077-1091, 2021.

[16] P. Singh, R. Srivastava, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, pp.1073-1080, 2021.

[17] S. Prasanth, M. R. Thanka, and V. Nagaraj, "Speech emotion recognition based on machine learning tactics and algorithms," *Materials Today: Proceedings,* vol. 17, pp. 632-646, 2021.

[18] H. A. Abdul Mohsin, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers & Electrical Engineering*, vol. 93, pp. 1071-1084, 2021.

[19] M. Ren, W. Nie, and Y. Su, "Multi-modal Correlated Network for emotion recognition in speech," *Visual Informatics*, vol. 3, pp. 150-155, 2019.

[20] C. Zheng, C. Wang, and N. Jia, "An ensemble model for multi-level speech emotion recognition," *Applied Sciences*, vol. 10, pp. 205-215, 2019.

[21] S. S. Poorna, and G. J. Nair, "Multistage classification scheme to enhance speech emotion recognition," *International Journal of Speech Technology,* vol. 22, pp. 327-340, 2019.

[22] J. C. Martinez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical signal processing and contro*l, vol. 55, pp. 1016-1028, 2020.

[23] M. Ali, F. Al Machot, and K. Kyamakya, "A globally generalized emotion recognition system involving different physiological signals," *Sensors*, vol. 18, pp. 1905-1920, 2018.

[24] F. Nasoz, K. Alvarez, C. L. Lisetti, N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cognition Technology,* vol. 6, pp. 4-14, 2004.

[25] D. Ayata, Y. Yaslan, M. E. Kamasak, "Emotion recognition from multimodalphysiological signals for emotion aware healthcare systems," *Journal of Medical and Biological Engineering,* vol. 40, pp. 149-157, 2020.

[26] F. Ren, Y. Dong, W. Wang, "Emotion recognition based on physiological signals using brain asymmetry index and echo state network," *Neural Computing and Applications*, vol. 31, pp. 4491-4501, 2019.

[27] X. Zhang, C. Xu, and M. Gao, "Emotion recognition based on multichannel physiological signals with comprehensive nonlinear processing," *Sensors*, vol. 18, pp. 3886-3898, 2018.

[28] Y. Fang, R. Rong, and J. Huang, "Hierarchical fusion of visual and physiological signals for emotion recognition," *Multidimensional Systems and Signal Processing,* vol. 32, pp. 1103-1121, 2021.

[29] S. Cunningham, H. Ridley, J. Weinel, R. Picking, "Supervised machine learning for audio emotion recognition. Personal and Ubiquitous Computing," vol. 25, pp. 637-650, 2021.

[30] E. Koh, and S. Dubnov, "Comparison and analysis of deep audio embeddings for music emotion recognition," *Arxiv Preprint*, vol. 21, pp. 6517-6530, 2021.

[31] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *Journal of Ambient Intelligence and Humanized Computing,* vol. 13, pp. 867-885, 2022.

[32] A. I. Middya, B. Nag, S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, pp. 1085-1098. 2022.

[33] S. N. Chennoor, M. Ali, T. K. Kumar, "Human emotion detection from audio and video signals. *Arxiv Preprint*, vol. 2, pp.1187-1190, 2020.

[34] A. P. Reddy, and V. Vijayarajan, "Audio compression with multi-algorithm fusion and its impact in speech emotion recognition," *International Journal of Speech Technology*, vol. 23, pp. 277-285, 2020.

[35] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals" *Computers in Human Behavior*, vol. 65, pp. 267-275, 2016.

[36] J. Chen, T. Ro, and Z. Zhu, "Emotion recognition with audio, video, EEG, and EMG," *IEEE Access*, vol. 10, pp. 13229-13242, 2022.

[37] J. Wei, X. Yang, and Y. Dong, "User-generated video emotion recognition based on key frames," *Multimedia Tools and Applications*, vol. 80, pp. 14343-14361, 2021.

[38] Y. R. Pandeya, B. Bhattarai, J. Lee, "Deep-learning-based multimodal emotion classification for music videos," *Sensors,* vol. 21, pp. 27-41, 2021.

[39] Z. Wang, X., Zhou, W. Wang, C. Liang, "Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 923-934, 2020.

[40] M. Singh, and Y. Fang, "Emotion recognition in audio and video using deep neural networks," *Arxiv Preprint,* vol. 20, pp. 129-141, 2020.

[41] D., Dresvyanskiy, and E. Ryumina, "An audio-video deep and transfer learning framework for multimodal emotion recognition in the wild,". *Arxiv Preprint*, vol. 20, pp. 567-579, 2020.

[42] S. Zhalehpour, Z. Akhtar, and E. Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image and Video Processing,* vol. 10, pp. 827-834, 2016.