International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

Real-Time Multimodal Emotion Recognition

1.Mrs.S. Bhargavi, 2.Mrs.B.Siva Kumari 3.Ms. M.Swathi, 4.Ms. K.Kalyani, 5.Ms.Ch.Rajasri

1. Faculty of Electronics and Communication Engineering, Bapatla women's Engineering College, Bapatla, Andhra Pradesh, India.
2. Faculty of Electronics and Communication Engineering, Bapatla women's Engineering College, Bapatla, Andhra Pradesh, India.
3. Student of Electronics and Communication Engineering, Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.
4. Student of Electronics and Communication Engineering, Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.
5. Student of Electronics and Communication Engineering, Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India.

ABSTRACT:

Safeguarding the well-being of women and children presents a challenging research endeavor. Multimodal emotion recognition poses a formidable task within this domain. The field of Human-Computer Interaction (HCI) heavily relies on multimodal data, encompassing audio, video, text, facial expressions, body motions, bio-signals, and physiological data, to predict the safety of women and children. Substantial research efforts have been dedicated to this cause. To develop an optimal multimodal model for emotion recognition, which integrates visual, textual, auditory, and video modalities, a novel deep learning framework is proposed. This framework involves a comprehensive analysis of data, feature extraction, and model-level fusion. Innovative feature extractor networks are tailored specifically for processing visual, textual, auditory, and video data. At the model level, an effective multimodal emotion recognition model is devised, synthesizing information from images, text, voice, and video. The proposed models exhibit impressive performance on three benchmark multimodal datasets, namely IEMOCAP, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Surrey Audio-Visual Expressed Emotion (SAVEE), achieving high predicted accuracies of 96%, 97%, and 97%, respectively. Comparative analysis with existing emotion recognition models further validates the efficacy and optimality of the proposed approach. The application of multimodal enhanced emotion recognition holds promise in predicting women and children's safety.

Index Terms: Facial Expression Recognition, Deep Learning, Multimodal, Women's Safety, Audio-Visual Media, Fusion.

INTRODUCTION:

Human Their emotional moods have a huge impact on communication, which is quite important. It can be difficult to determine the emotional states in a variety of areas with a wide range of applications, including lie detection, audio- visual police work, affectional computing, online teaching and learning, online meetings, human-computer interface (HCI),and many ,many more. Thein vestigation of emotional variability is an essential component of psychological, mental adaptability ,and well-being research. Additionally, machines need to be able to identify human emotions in order to make better judgements. The advancement of techniques to let intelligent computers accurately assess users' emotions is now essential for the progress of civilizations in cethey must become an integral part of our daily lives Research commonly

ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

uses two models:dimensionalmodels 1ResearchScholar,Dr.BabasahebAmbedkarTechnological University, Lonere (India), nandawagh@ dbatu.ac.in 2ProfessorandHeadofDepartment,Dr.BabasahebAmbedkar Technological University ,Lonere(India),srsutar@dbatu.ac.in and discrete feeling models, despite the fact that there are numerous emotional models in the literature [4,5]. The former describes emotions as an infinite spectrum whereas the latter portrays them as unique values. Each model is regularly used by psychologists to evaluate emotions. Using voice data, physiological data, facial expressions, body language, and a variety of other elements, it is common practise to determine a person's moods. Each of those modalities hasits own unique characteristics, thus combining them yields a chicde paction of options that can swiftly play- act feeling recognition. Wishing on a single modality for feeling recognition is futile, as has been shown in past experiments [6,][7]. Particularly, current research shows that using many modalities (audio, video, text, etc.) for emotion recognition yields considerably better results than using only one[8].The former refers to emotions as a never-ending spectrum, where as the latter shows them as discrete values. Each model is widely used by psychologists to investigate emotions. Multimodal period analysis of emotional states has received a great deal of attention recently from people all around the world under the a fore mentioned paradigm .Dynamic multimodal analysis of human emotions outperforms static analysis of human emotions because it takes into consideration elements like fluctuations in eye movements and facial expressions over time as well as voice characteristics [9]. Finding a cost-effective multimodal model in this field of research that is also not unduly complicated or computationally taxing is still a challenge. Therefore ,it is an ticipated that this work will be able to detect emotional states using a multimodal fusion model that considers both audio and video data. It is reasonably lightweight and reasonably priced. The two methods that have drawn the most interest in the last 10 years are multimodal emotion recognition (MER) and facial expression recognition(FER).FER will be produced using a variety of techniques, including face feature analysis, bi-signals, and the tried-and-true multi-modal approach [6],[3]. The various informational forms show the acknowledged outperformed lead. The modalities encompass all types of information, such as text cues, face images, acoustic expressions, linguistic data, semantic patterns, physical movements, eye gaze patterns, gestures, and electroencephalography (EEG) signals [2], [36], and [36] signals, as well as semantics like these signals. Numerous style strategies have been presented in recent research to automatically recognise affectional outcomes such as valence, arousal, dominance, and other emotion types [36, 37]. Emotions are the primary social communication components of importance. At some time in the future, depending on how it displays itself, our eyes will be able to tell whether it is okay or not. Facial Features Recognition (FER) is a simpler problem. However, those who are deaf, visually disabled, or less sensitive are unable to convey their own sentiments. The most difficult FER evaluations are those performed by machines. By using the proper collection of algorithms and employing a multimodal strategy, the right set of emotions can be predicted. The necessity for girls' safety has increased, as it does in modern settings, and technology is crucial in modifying it. People often think of the face as their mental hub. As it makes different facial expressions, the face will produce a range of small Signals [56].

The term "emotional sensation" refers to the mental state that the human mind is capable of seeing and categorizing in order to appraise a range of appearances Interaction becomes more personalized and distinctive when computer algorithms attempt to grasp the user's purpose. The call from the computer becomes more significant as it starts to ask questions based on the user's mood. An autonomous vehicle might decide to travel further if the user's emotions are prone to anger, for instance. Itisan excellent and extraordinarily sensitive indicator for figuring out how people behave, intend, think, and feel. With human safety and the aid of a feeling index, we frequently carryout autonomous tasks like marketing, observation, car safety, and appraisal .It will be simpler for people to communicate with technology in all imaginable

ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

ways thanks to the human- computer interface(HCI)industry. Interaction becomes more personalized and distinctive when computer algorithms attempt to grasp the user's purpose. The call from the machine becomes more important as it begins to pose queries based on the user's emotional state. An autonomous vehicle might decide to travel further if the user's emotions are prone to anger, for instance. It should start playing calming music when auser is feeling down. Young girls and kids can receive rapid support with pressing problems during the victimization phase thanks to the multimodal feeling identification technology, which assesses sentiments efficiently. Having outstanding, accurate, and acceptable judgement is crucial since good judgement is the capacity to discern between various emotions. A person's spirit can be effectively communicated through their face, voice, body language, gestures, movies, and even particular circumstances. Research of researchers shows that these method attempt to accurately predict emotions to varied degrees.

To obtain higher precision, a variety of techniques and tactics have been discovered and are currently being applied worldwide. Paul Ekman, an associate degree creator,claimed that happiness, anger, surprise, sorrow, concern, and dislike are the six basic emotions that comprise a human emotional state. Ekman also noted the propensity to obfuscate human emotions by utilizing units of action (AU) [9]. play [1]. Numerous additional modalities eventually emerged as a result of the fact that facial features predominated among the modalitiesforidentifying moods. The major objective of this paperistodo multimodal emotion recognitionutilizing text, facial images, and sound. The video is immediately used as the source for the multi modalities. Girls and children are the targets .Violence is contagious in the society that is governed by men. It's too late to stop the breakout scenario now. How can crime against girls and children be addressed and reduced? By implementing positive society improvements, crime can be decreased and perhap seven managed to some extent. Mole station,robbery, and rape incidents happen frequently. Physically abusing girls and children as part of domestic violence is common. We treat every square foot like an animal. Most of the time, while counting, the simplest square measurement isused.The "WOMEN EMPOWERMENT" world is still a "fantasy" on a grand scale. Depressive disorders are more prevalent in girls and young people. Byusingdeeplearningandmachinelearningtorecognise multimodale motions, the problem of women's and children's safety will be resolved. The technology paradigm will lessen the victimization of women and children .Concerns about the safety and wellbeing of women and children, particularly in public settings, have grown over the past several years. The necessity for rigorous protocols to secure their protectionand give prompt aid has been underlined by incidents of harassment,assault,andabuse.In this situation,multimodal emotion identification technology has emerged as a possible solution. This technology uses a variety of sensory inputs, including body language, vocal intonation, and facial expressions, to precisely analyze and decode human emotions. Multimodal emotion identification has the potential to considerably improve current safety precautions for women and children while also enabling people to react skillfully to looming risks. Thepurpose ofthisresearch paperisto study the idea of multimodal emotion identification and its applications in thecontext of women's and children's safety, emphasizing the importance of this idea in establishing safer environments for those who are more vulnerable.The goal of this project is to provide a way for employing convolution neural networks to recognize emotions related to the protection of girls,women,and children.When employing a digital camera to capture people's faces over the given time period, the model is utilized to identify their moods. Human emotions significantly influence communication and are crucial across various domains such as lie detection, police work, and online interactions. Researchers commonly use two models to evaluate emotions: dimensional and discrete feeling models, each with its own advantages. Multimodal analysis, combining audio, video, and text, has shown better results in emotion recognition compared to single-modal

ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

approaches. Facial expression recognition (FER) and multimodal emotion recognition (MER) are two prominent methods for emotion analysis. Multimodal approaches consider various sources of information, including facial expressions, voice data, and physiological signals. These methods have practical implications, especially in enhancing safety measures for vulnerable populations like women and children. By employing technologies like deep learning and machine learning, it becomes possible to predict and mitigate risks related to emotional states. Overall, the aim is to develop effective strategies for emotion recognition to create safer environments, particularly for those at risk.

The research contribution is predictive analysis of emotion based on multimodalities and utilized for women and children safety.

1. The proposed bust convolution neural network model building and decision level fusion for high accuracy.
2. Deep learning-based feature extractor networks for images, text, video, and audio.
3. A model-level fusion of the video associate degreed audio options is performed to make an optimum multi modal feeling recognition mode

The structure of part of this article is as follows: These cond part presents the literature review, the challenges of multimodal emotion recognition and multimodal data sets. The third part focuses on transfer learning, hyper tuning ,and novel convolution neural network model design with optimized parameters. The fourth evaluates the relevant experimental results and analyzes. The fifth part state summary, the conclusion and future scope.

2. METHODOLOGY

*1.1.*Datasets MultimodalDatasets

IEMOCAP, SAVEE and RAVDESS are multimodal datasets used for experimentation.

Three benchmark multimodal databases, the IEMOCAP [3], SAVEE [2] and the RAVDESS [1] are used in this

IEMOCAP has 12 hours of audiovisual material on 10 actors, including discussion between an actor and an actress that was both scripted and spontaneously recorded [1],[3]. The audiovisual data is compiled into short sentences that last between three and fifteen seconds, which are subsequently identified by the assessors. Three to four different people analyseeach statement.Tenoptions(neutral,happy,sad,angry, surprised, afraid, disgusted, frustrated, delighted, other) were provided on the evaluation form.Weonlyexaminefourofthem becausethatiswhatpriorresearchhasdone:anger,excitement (happiness), neutrality, and sadness. Very Good Form According to earlier studies, we consider emotions when at least two experts concur with their choice,1],[10][15].

The Surrey Audio-Visual Expressed Emotion (SAVEE) data base has been documente dasa requirement for the creation of an automatic emotion recognition system. The database contains 480 British English utterances recorded from 4 male actors portraying 7 different emotions. The sentences were phonetically balanced for each mood and taken from the typical TIMIT corpus. High-end audio-visual equipment was used to record, process, and label the data in a visual media lab. Ten volunteer sex amined there cording sunder auditory,visual,and audio-visual circumstances in order to assess the performance quality. For the auditory, visual, and audio-visual modalities, classificationsystemswerecreatedusingstandardfeaturesand classifiers, and speaker-independent identification rates of 61%, 65%, and 84% were obtained, respectively [2].

A multimodal dataset called RAVDESS [1][11] contains 7356 files totaling24.8GB.These statistics are from24professional actors (12 men and 12 women), who each spoke two lexically similar phrases with a North American accent. Disgust, surprise,fear,sadness,joy,anger,andcalmemotionscanallbe heard in speech. Every display of emotion has two unique emotionalintensitylevels(strongandnormal),aswellas neutralexpression.

ThedatamodalitiesincludedinthisdatabaseareAudio-Video, Audio-Only, and Video-Only (without sound).

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

Because the objective of this work is to perform multimodal emotion recognition, which requires the use of both audio and visual dataforeachactor,afractionofthevideospeechfiles(i.e.,files with both audio and visual modalities) are employed.

There are 1440 files in total, which are divided into eight emotion classes: fearful, disgust, angry, sad, happy, calm, neutral, and surprised.

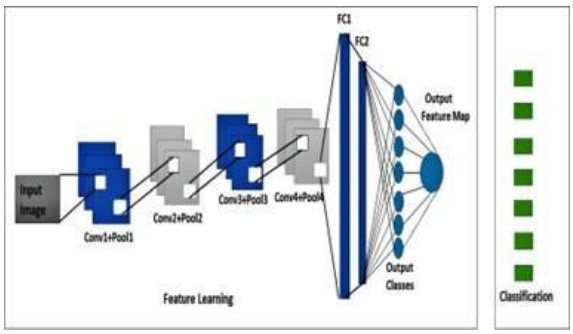
Four researchers and students from the University of Surrey, ranging in age from27to31,recorded the data for the SAVEE [2][10]dataset.This data set contains480 audio–visualfiles, with 120 utterances for each speaker.Alltheaudio–visualfiles arein.avid format and there are seven emotion classes namely surprise, sadness, neutral, happiness, fear, disgust, and anger. FER2013database.The data collectionused for the application wastheFER2013datasetfromtheKagglechallengeon FER2013[38]. The database is used to incorporate the Facial Expression detection frame work.The data set consists of35,887 pictures, splitinto3589experiments and28,709images of trains.The data set includes an other3589privatetestimagesfor the final test.

Multimodal approach uses text ,images,speech,and video sequences.

Enhancement for the anticipated CNN. A fine-tuning methodology replaces the pre-trained model's completely connectedlayerswithanewsetofcompletelyconnectedlayers when training a model on a given dataset. It also uses backpropagation to fine-tune all or some of the kernels in the pre-trained convolutional sub caste base.The active factors that control the affair size of the convolutional subcaste include padding,stride,batch size,sludge,sliding window,and literacy rate parameter. Padding is necessary to give the input border bottoms. The height and range parameters are distributed by Stride. Small stride lengths provide wide fields and massive, widely over lapping affairs.

The open fields lap lower, with longer strides and less limitation. While fine-tuning all of the layers in the convolutional base,it is possible to tunea considerable portion ofthedeeperlayers.Theproposed model in this work has four layers of complexity and two fully connected layers.In order to complete this task, just the completely linked layers that serve as a classifier and the detailed point block material at high position would need to be trained. Since we only had 7 sensations, which is a disparity,the Soft Max rating was reset to 7 grades from 1000 species. an anticipated CNN pipeline.

A net workthatrecyclesinputusing subcasting. The latter two levels are entirely connected to the last four layers of complexity and new pooling. Batch normalization, ReLU subcaste, and a completely linked subcaste of each of the four network topologies are used to handle any complexity.The new thick subcaste is used at the end of the four complication layers, which are connected to the two totally connected layers. One of the two possibilities serves as the foundation for the general channel in the proposed CNN model. Fig. 3. The basic mechanism for recognizing emotions is depicted in Fig. 4. On the face picture or modality, preprocessing and feature extraction are carried out. It is anticipated that human beings would be classified according to their performed positive and negative mood states.



ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

Fig. ProposedConvolutionNetworkArchitecture

Data preprocessing and feature engineering

In this section we describe how we extracted our video and audio features and prepared our multimodal labeled datase

MultimodalFacialEmotionRecognition

The proposed research focuses on utilizing a negative set of emotions, including anger, sadness, disgust, and fear, to predict safe and unsafe situations for the safety of women and children. This involves a multimodal approach to emotion detection, incorporating audio, video, text, and facial expression recognition, which has shown to outperform uni-modal approaches. The most effective classifier combines text, audio, and visual information for improved emotion detection accuracy.

Deep learning techniques, such as MT-CNN and Convolutional Neural Networks with Hyper Parameters, are utilized for image emotion detection, while Transfer Learning and hyper-parameter tuning are employed for heightened face emotion recognition. The suggested method achieves a recognition rate of 95.91% on training data, outperforming sophisticated approaches.

Additionally, a four-layer Convolutional Neural Network model is proposed for real-time facial expression recognition, integrating CNN, Local Binary Pattern (LBP) features, and Oriented FAST and rotating BRIEF (ORB) techniques. The research emphasizes the importance of using aggregated photos from various databases to enhance generalization and prediction precision.

Overall, the research aims to advance emotion detection and prediction methodologies, particularly for ensuring the safety of women and children, through innovative multimodal approaches and deep learning techniques. The ConvNet architecture demonstrates outstanding performance across large, medium, and small datasets. Training accuracy surpasses 97% within a few epochs, indicating excellent calibration of the model. Moreover, the classification accuracy on various datasets, including IEMOCAP, SAVEE, and RAVDESS, ranges from 96% to 97%. This research contributes to predictive emotion analysis through multimodal approaches, particularly for ensuring the safety of women and children. The proposed robust convolutional neural network model, coupled with decision-level fusion, achieves high accuracy.

Datapreprocessingandfeatureengineering

In this section we describe how we extracted our video and audio features and prepared our multimodal labeled dataset In video preprocessing and feature engineering, each video recording is segmented into six frames per three-second clip to capture spatiotemporal information related to emotions. OpenCV, a machine vision software, is used to extract frames, and a Caffe-based facial detector is employed to identify facial landmarks. The frame size is reduced by 40%, and the blob From Image() function is applied for image preparation, including normalization and mean pixel intensity deduction. Faces are detected using discovery scores, and weak detections are filtered out based on a confidence threshold. The facial region is cropped and resized to 64 by 64 pixels for feature extraction.

Audiopreprocessingandfeatureengineering

Every video has its audio material extracted using them oviepy Python package. Since the retrieved audio was discovered to be a stereo file,it was split into a monofile using the pydub module [55]. then the mono audio file is used for feature extraction. Numerous audio features are retrieved and kept in the audio feature list, in particular MFCC, melspec- trogram, spectral difference, and tonnetz. Below is a list of the significance of each of those choices..

TransferLearning,MultimodalFusion

Akey ideain deep learning istransferlearning. It employsthe ideas of reuse and makes use of models that havebeen trained to tackle one problem a sajumping off point for an other related issue. The most adaptable

ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

method of learning is transfer learning, which enables previously trained models to be utilised directly as feature extraction preprocessing and merged into whole new models[14][18][15].ManypotentImageNetimage recognition task models, including VGG, Inception, and ResNet, are easily accessible using Kera's. The pre-trained traditional models for image categorization are ImageNet MobileNet, MobileNetV2, Xception, VGG16, VGG19, ResNet50, ResNet101, ResNet101V2, ResNet152V2, InceptionV3, InceptionResNetV2, and DenseNet. As seen above, ImageNet's annual Large Scale Visual Recognition Challenge (ILSVRC) featured a variety of high-performance picture categorization algorithms. Due to the image source utilised in the competition,this task is frequently referred to as simply "ImageNet," and it led to a number of advancements in convolutional neural network architecture and training. The foundation for transfer learning in machine vision applications can be efficiently employed withany model. Therearealot of admirable uses for it, including

Features that were learned and are useful As they were trained onmorethan1,000,000photographs over1,000categories,the models learned to recognise the general properties of the images. Modern Performance: The models showed modern performance and excelled in the particular image emotion detection job for which they were created.Simple to Access: ManylibrariesofferstraightforwardAPIsfordownloadingand using models, and model weights are available as free downloads. Several alternative deep learning packages, including Keras[37][39],allow for the down loading and usage of model weights in the same model architecture.4. Experimentation and results

This section presents the experimental setup and result analysis for the approaches taken to train multimodal emotion detection. Following are the approaches:

1. For training model using CNN, for IEMOCAP Dataset, the experiment results are as below:

Epoch:200
Batch: 228 TrainingResults Epoch 200/200
228/228 [==============================] 18s
80ms/step loss:0.1391
accuracy:0.9691
val_loss:1.7726
val_accuracy: 0.6663.

ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

3.CONCLUSION:

In this research article, a multimodal facial emotion recognition system is extensively investigated, and a multimodal convolutional network is proposed. The research emphasizes transfer learning, hyper parameter tuning, and the facial emotion recognition learning process. We have used the transfer learning concept of the popular current neural convolution algorithm combined with Mobile Net50,which has had a worthy performance for effect on the multi-class classification. While performing validation on the data set, the experimental evaluation result has a good exactness and a good recognition effectin terms of averagere cognition precision.In future research, we will focus on exploring diverse facial emotion detection and will try to collect more real time emotionalimages,video,andtranscriptsthaninthisexperiment in order to optimize and suggesta better algorithm to train the hyper parameters of the multilayer feedback neural network, such as weights and bias. We will also evaluate optimization algorithms based on the previous method to increase the performance of a multilayer feed forward neural network. We will continue to search for shapes based on a deep residual network to improve the accuracy of facial expression recognition. Todevelop robustdeep learning model foremotion recognition of physically challenged,blind, deaf, andmentally retired human beings and old ageperson is future work of this research.

4.Acknowledgment:

We are extremely thankful to Dr.Abhay E.Wagh,Director, Department of Technical Education, Government of Maharashtra; we are most deeply thanks to Dr. Santosh T. Yadav for his consistent motivation and lifetime support for research. We are very muchthankfulto Dr.Sanjay Nalbalwar, Head of the Department of Electronics and Telecommunication, Dr. Babasaheb Ambedkar Technological University Lonere. We thanks Dr. Brijesh Iyer, Assistant Professor, Department Electronics and Telecommunication, DBATU Lonere. We thanks a lot Dr. D.R. Nandanwar, Principal, Government Polytechnic Awasari for his moral and coherentsupport.Mr.Vishal Supekar for his unconditional and great support forever. We thanks Dr.Vijay Kohle and Mrs. Seema Kolhe fortheir worthy and inspirational supportat step of life. Thanks Dr.S.M.Jadhav,Dr. Kiwalekar, Dr. L.D.Netak and Dr. V.J. Kadam for their significant guidance and consistent support.I thank to former head of Department Mr.U.V.Kokate,Government Polytechnic,Pune for his guidance.I thanks Mrs. J.R.Hange, Head of Department, Computer Engineering,Government Polytechnic Awasari for her support. I thank sall my colleagues form GovernmentPolytechnic Pune and Government Polytechnic Awasari.

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

5.RESULTS AND DISCUSSION:

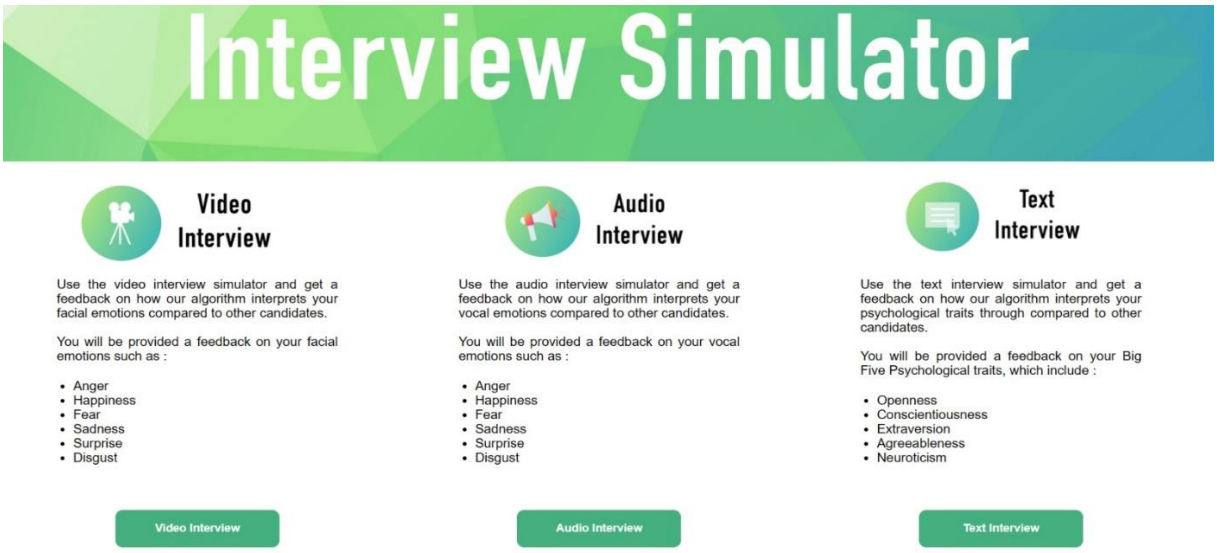


Fig.5.1 App Simulator

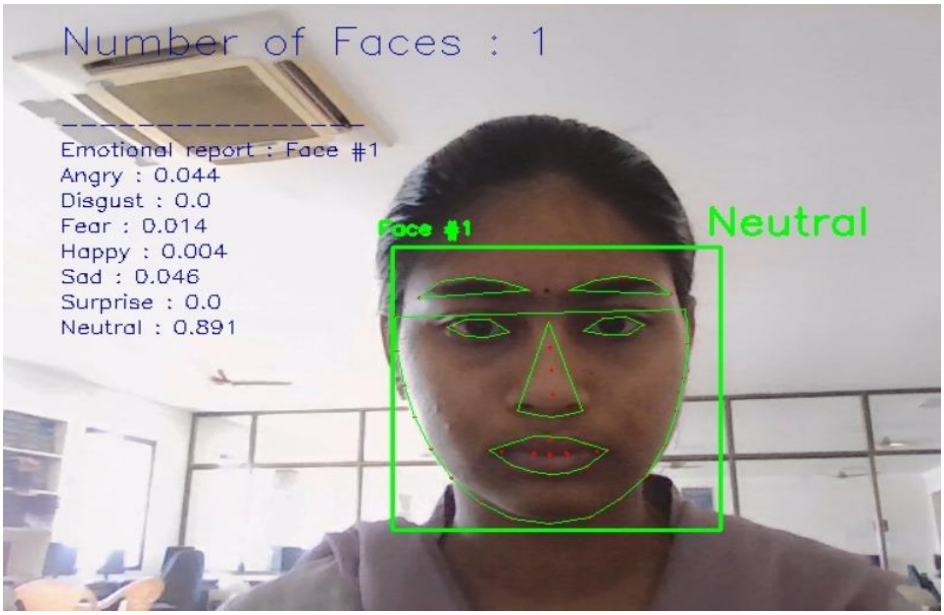


Fig.5.2 Vedio Output

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

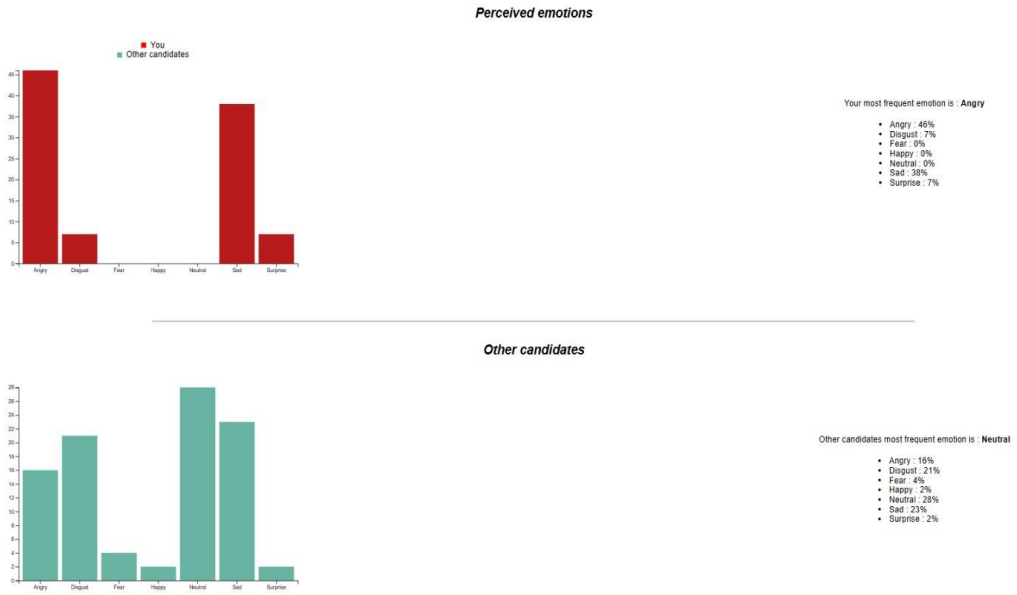


Fig.5.3 Audio Output

ISSN: 2583-6129
DOI: 10.55041/ISJEM01414

6.REFERENCES:

Puri,T.,Soni,M.,Dhiman,G.,IbrahimKhalaf,
O. and Raza Khan, I., 2022. Detection of emotion of speech for RAVDESS audio using hybrid convolution neural network. Journal of Healthcare Engineering,

[1] Singh, P., Srivastava, R., Rana, K.P.S. and Kumar,V.,2021.Amultimodalhierarchicalapproachto speechemotionrecognitionfromaudioandtext. Knowledge-Based Systems, 229, p.107316.

[2] Busso,C.,Bulut,M.,Lee,C.C.,Kazemzadeh,
A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42, pp.335-359.

[3] P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari,etal.,Combiningmodal-ityspecificdeepneural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 543–550.

[4] N. Srivastava, R. Salakhutdinov, et al.,MultimodallearningwithdeepBoltzmannmachines,in:NIPS, Vol. 1, Citeseer, 2012, p. 2.

[5] J.Ngiam,A.Khosla,M.Kim,J.Nam,H.Lee,
A. Ng, Multimodal deep learn-ing, in: InternationalConference on Machine Learning (ICML), Bellevue,WA, 2011, pp. 689–696.

[6] Y. Wang, L. Guan, Recognizing humanemotional state from audiovisualsignals, IEEE Trans.Multimed. 10 (5) (2008) 936–946.

[7] C.Busso,Z.Deng,S.Yildirim,M.Bulut,C.M. Lee, A. Kazemzadeh, S. Lee,

[8] U. Neumann, S. Narayanan, Analysis of emotionrecognitionusingfacialexpressions,speechand multimodal information, in: Proceedings of the 6th International Conference on Multimodal Interfaces, 2004, pp. 205–211.

Y. Yoshitomi, S.-I. Kim, T. Kawano, T.Kilazoe, Effect of sensor fusionfor recognition ofemotional states using voice, face imageandthermalimage of face, in: Proceedings 9th IEEE InternationalWorkshop on Robotand Human InteractiveCommunication. IEEE RO-MAN 2000 (Cat. No.00TH8499), IEEE, 2000, pp. 178–183.