Multimodal Emotion Recognition Using Multiple AI Algorithms

Inukonda Manikanta
PG scholar, Department of MCA, DNR College, Bhimavaram, Andhra Pradesh.
B.S.Murthy
(Assistant Professor), Master of Computer Applications, DNR college, Bhimavaram, Andhra Pradesh.

Abstract: This project aims to develop a real-time multimodal emotion recognition system that detects emotions from video, speech, and text inputs. The system activates the camera to identify emotions instantaneously. To achieve this, we train a robust model using four state-of-the-art deep learning algorithms: VGG19, a deep convolutional neural network known for its intricate feature capture; ResNet50, a 50-layer network that overcomes the vanishing gradient problem; MobileNetV2, a lightweight model optimized for mobile and edge devices; and Xception, which uses depth-wise separable convolutions for high performance. The project involves comparing the accuracy of these algorithms to determine the most effective approach for real-time emotion recognition.

1. Introduction

his chapter aims to explore innovative methods for enhancing human-computer interaction by making computers more aware of the user's emotional and attentional expressions.[1] The aim of this study is to develop a robust and comprehensive multimodal emotion recognition system that integrates various types of digital data—text, speech, and images—to accurately detect and interpret human emotions. By leveraging the strengths of multiple advanced AI algorithms, including Xception, MobileNetV2, ResNet50, and VGG19.

The project, "Multi-Modal Emotion Recognition," aims to develop a real-time system that activates the camera to identify emotions instantaneously. This will be achieved by training a robust model using four state-of-the-art deep learning algorithms: VGG19, known for its intricate feature capture; ResNet50, a 50-layer network that addresses the vanishing gradient problem; MobileNetV2, optimized for mobile and edge devices; and Xception, which utilizes depth-wise separable convolutions for high performance.

The project involves comparing the accuracy of these algorithms to determine the most effective approach for real-time emotion recognition. The model should be designed to operate in real-time by turning on the webcam and also processing images, speech, text, and video files when provided through a file path or upload.

After training and comparing the four models, the project will use the most accurate model along with MobileNetV2 for further processing. This ensures the system leverages both the best-performing model and a lightweight model optimized for mobile and edge deployment.

This chapter provides an overview of research efforts focused on classifying emotions using various modalities, including audio, visual, and a combination of both audio-visual. [2] Multimodal Emotion Recognition refers to the process of detecting and interpreting human emotions through the integration of multiple types of data or modalities. Unlike traditional approaches that rely on a single type of data, such as facial expressions or speech, multimodal emotion recognition combines various sources of information, including text, speech, and images, to achieve a more comprehensive and accurate understanding of emotional states.

1.4.1 Key Modalities in Multimodal Emotion Recognition:

1. **Text:** Analysing the sentiment and emotional content in written language. This can include social media posts, emails, or transcriptions of spoken language. Techniques such as Natural Language Processing (NLP) are used to extract emotional cues from text.
2. **Speech:** Evaluating vocal characteristics such as tone, pitch, and intensity. Speech

emotion recognition can identify emotions by examining how something is said rather than what is said. Acoustic features and prosody are important in this analysis.

3. **Images:** Interpreting facial expressions, body language, and other visual cues. Computer vision techniques are used to detect and classify emotions based on facial movements and expressions.

This survey seeks to bridge the gap by offering a comprehensive overview of recent advancements in multimodal emotion recognition (MER) based on deep learning. [4]

Traditional emotion recognition systems often rely on a single type of data, such as text, speech, or images, which can limit their accuracy and robustness. To overcome these limitations, our project focuses on developing a multimodal application that integrates various types of digital data—text, speech, and images—to detect emotions more comprehensively.

This paper presents a deep dual recurrent encoder model that simultaneously analyses both text and audio data to provide a deeper and more complete understanding of speech information. [5]

II. LITERATRURE SURVEY

Recent technological advances have expanded human-computer interactions beyond traditional methods, introducing modalities like voice, gesture, and force-feedback. However, integrating emotion recognition remains a key challenge for natural interaction. Emotions are crucial in human communication, and enabling computers to understand them is desirable for various applications. This chapter reviews fundamental research and recent developments in emotion recognition from facial, voice, and physiological signals, addressing them independently. It also discusses the complexities of multimodal emotion recognition and the use of probabilistic graphical models for integrating different modalities, as well as challenges in obtaining reliable affective data and ground truth. [1]

Recent advancements in human-computer interaction aim to enhance the naturalness and user-friendliness of interactions by recognizing users' emotions. This chapter reviews research on emotion classification through audio, visual, and combined audio-visual modalities. It begins with theories of emotion to define categories and describes the creation of relevant databases. The authors outline fifteen datasets and the features used to represent emotional content, discussing methods for feature selection and reduction to remove noise. They highlight popular classifiers for determining emotion classes and methods for fusing multimodal information. The chapter concludes with suggestions for future research directions in emotion recognition. [2]

New research in human-computer interaction focuses on incorporating users' emotional states to enhance interface seamlessness, applicable in diverse fields like education and medicine. Emotional recognition techniques neuroimaging strategies, physiological signs, facial images, and encompass expressions. This paper reviews deep learning-based emotional recognition of multimodal signals, comparing their applications with current studies. It highlights those multimodal affective computing systems, compared to unimodal solutions, offer higher classification accuracy. This accuracy depends on the number of emotions, features extracted, classification systems, and database consistency. The paper also discusses theories on emotional detection methodologies and recent advances in emotional science, aiming to improve the understanding of physiological signals and emotional awareness issues. [3]

Multimodal emotion recognition (MER) combines signals like text, speech, and facial cues to identify human emotional states, playing a crucial role in human-computer interaction (HCI). Recent advancements in deep learning and multimodal datasets have led to significant developments in MER. However, comprehensive reviews of these achievements are lacking. This survey addresses this gap by analyzing current multimodal datasets, methods for emotional feature extraction, and various MER algorithms. It particularly focuses on model-agnostic fusion methods and fusion techniques in deep models, offering guidance for researchers in this dynamic field. [4]

Speech emotion recognition is a challenging task, typically relying on audio features to build effective classifiers. This paper introduces a novel deep dual recurrent encoder model that simultaneously utilizes text and audio data for improved speech emotion recognition. By encoding information from both audio and text sequences using dual RNNs and combining these sources, the model captures emotional dialogue more comprehensively. Extensive experiments demonstrate that this approach outperforms previous state-of-the-art methods on the IEMOCAP dataset, achieving accuracies between 68.8% and 71.8% for classifying emotions into angry, happy, sad, and neutral categories. [5]

Multimodal emotion recognition has garnered increasing interest due to its potential for enhanced performance by leveraging diverse information sources. This work explores using images, text, and tags for emotion recognition, addressing the often-ignored challenge of "missing modality" where social media posts may lack one or more types of content. To tackle this, we propose a multimodal model within a multitask framework that can train with any combination of modalities and predict emotions even when some modalities are missing. Our approach proves robust against missing modalities at test time and allows fine-tuning with unimodal and bimodal data to enhance performance. Experiments demonstrate that multitask learning also acts as a regularization mechanism, improving generalization. [6]

Emotions are key to human communication, and research on emotion recognition has grown significantly. Recently, multimodal emotion recognition, combining text, video, speech, and physiological signals, has gained importance for its ability to improve recognition rates. This paper pre-processes text, video, and speech, data from the IEMOCAP dataset, utilizes deep learning neural networks for feature extraction, and performs feature-level information fusion. Focusing on five emotions—happy, neutral, sad, excited, and angry —the proposed model achieves a validation accuracy of 0.68383 and a training accuracy of 0.9541 improving speech emotion recognition accuracy by 0.11751. [7]

III. PROPOSED METHOD

3.1 Proposed work

In this project we are developing multi-model application to detect emotion from various types of digital data such as Text, Speech and Images. To detect emotion we have trained and evaluate performance of multiple AI algorithms such as Xception, MobileNetV2, ResNet50 and VGG19. We evaluate each algorithm's performance in terms of accuracy, precision, recall, and FSCORE.

All training datasets are downloaded from KAGGLE repository by using terms like Speech Emotion, Text Emotion and facial expression images.

To implement this project, we have designed the following modules,

1) **Upload Emotion Dataset:** Using this module, we can load datasets into applications.
2) **Pre-process Dataset:** using this module we will apply processing techniques such as normalization, shuffling and splitting dataset into train and test and split ratio for training is 80% and testing ratio is 20%
3) **Train Multimodal Images:** using this module we will train and load all 4 algorithms and then perform prediction on test images data to calculate accuracy and other metrics
4) **Train Multimodal Speech:** Using this module, we will train and load all four algorithms, then perform predictions on test speech data to calculate accuracy and other metrics,
5) **Train Multimodal Text:** Using this module, we will train and load all four algorithms, then perform predictions on test text data to calculate accuracy and other metrics,
6) **Accuracy Comparison Graph:** will plot comparison graph between all algorithm performance
7) **Predict Emotion from Webcam**: this module will open a webcam and then prediction emotion using live faces
8) **Predict Speech Emotion:** using this module will upload speech audio file and then best performing model will be applied to predict emotion from speech audio
9) **Predict Text Emotion:** using this module will upload text file with sentences and

then best performing model will prediction emotion from TEXT.

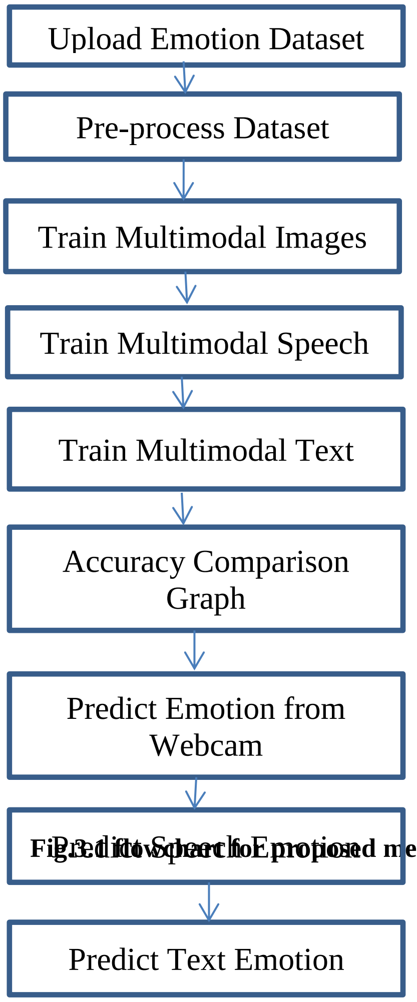


Fig.3.1 flowchart for proposed method

To run project double click on 'run.bat' file to get below screen

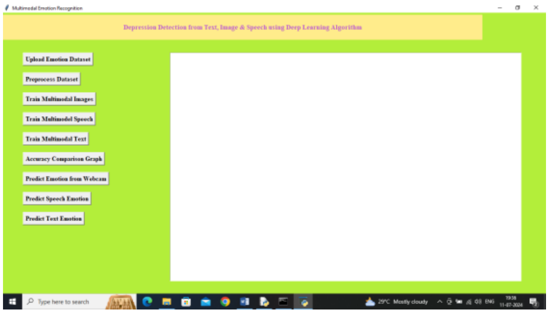

Multimodal Emotion Recognition
Depression Detection from Text, Image & Speech using Deep Learning Algorithm
Upload Emotion Dataset
Preprocess Dataset
Train Multimodal Images
Train Multimodal Speech
Train Multimodal Text
Accuracy Comparison Graph
Predict Emotion from Webcam
Predict Speech Emotion
Predict Text Emotion

Fig.4.1 'run.bat' file

In above screen click on 'Upload Emotion Dataset' button to load dataset and get below page

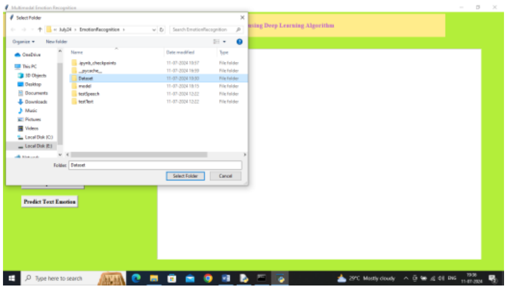
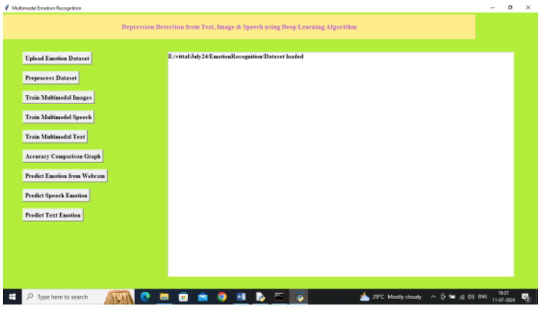


Fig.4.2 selecting and uploading entire 'Dataset' folder

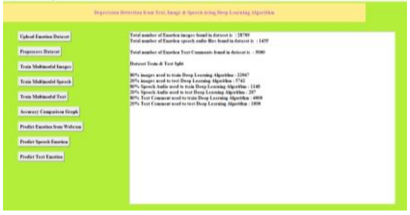
In above screen selecting and uploading entire 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below page


Multimodal Emotion Recognition
Depression Detection from Text, Image & Speech using Deep Learning Algorithm
Upload Emotion Dataset
Preprocess Dataset
Train Multimodal Images
Train Multimodal Speech
Train Multimodal Text
Accuracy Comparison Graph
Predict Emotion from Webcam
Predict Speech Emotion
Predict Text Emotion

```
E:/vittal/July24/EmotionRecognition/Dataset loaded
```

Fig.4.3 dataset loaded

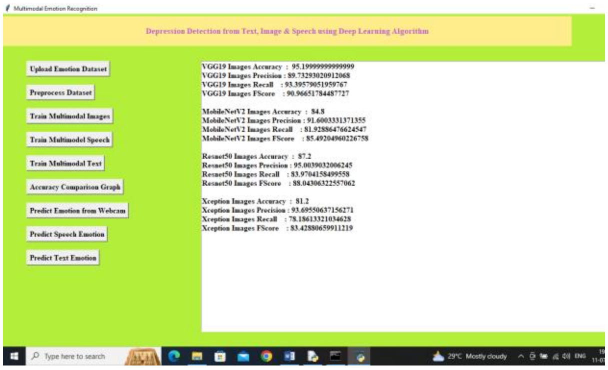
In above screen dataset loaded and now click on 'Pre-process Dataset' button to process dataset and get below page



```
Total number of Emotion images found in dataset is : 28709
Total number of Emotion speech audio files found in dataset is : 1435

Total number of Emotion Text Comments found in dataset is : 5080

Dataset Train & Test Split

80% images used to train Deep Learning Algorithm : 22967
20% images used to test Deep Learning Algorithm : 5742
80% Speech Audio used to train Deep Learning Algorithm : 1148
20% Speech Audio used to test Deep Learning Algorithm : 287
80% Text Comment used to train Deep Learning Algorithm : 4064
20% Text Comment used to test Deep Learning Algorithm : 1016
```

Fig.4.4 dataset size of each format

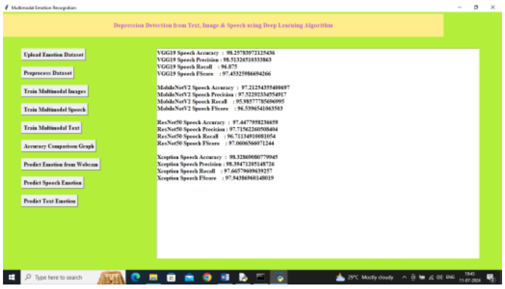
In above screen can see dataset size of each format and then can see train and test size of each dataset and now click on 'Train Multimodal Images' button to train all 4 algorithms on facial images and get below page


Multimodal Emotion Recognition
Depression Detection from Text, Image & Speech using Deep Learning Algorithm
Upload Emotion Dataset
Preprocess Dataset
Train Multimodal Images
Train Multimodel Speech
Train Multimodal Text
Accuracy Comparison Graph
Predict Emotion from Webcam
Predict Speech Emotion
Predict Text Emotion

```
VGG19 Images Accuracy  :  95.19999999999999
VGG19 Images Precision : 89.73293020912068
VGG19 Images Recall    : 93.39579051959767
VGG19 Images FScore    : 90.96651784487727

MobileNetV2 Images Accuracy  :  84.8
MobileNetV2 Images Precision : 91.6003331371355
MobileNetV2 Images Recall    : 81.92886476624547
MobileNetV2 Images FScore    : 85.49204960226758

Resnet50 Images Accuracy  :  87.2
Resnet50 Images Precision : 95.0039032006245
Resnet50 Images Recall    : 83.9704158499558
Resnet50 Images FScore    : 88.04306322557062

Xception Images Accuracy  :  81.2
Xception Images Precision : 93.69550637156271
Xception Images Recall    : 78.18613321034628
Xception Images FScore    : 83.42880659911219
```

Fig.4.5 performance of all algorithms on face expression

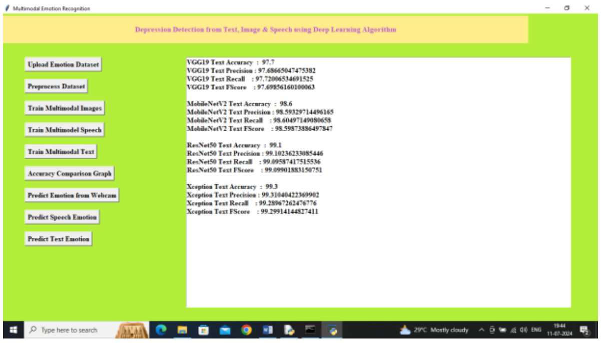
In above screen can see performance of all algorithm on face expression images and now click on 'Train Multimodal Speech' button to train all algorithms on speech data and get below page


Multimodal Emotion Recognition
Depression Detection from Text, Image & Speech using Deep Learning Algorithm
Upload Emotion Dataset
Preprocess Dataset
Train Multimodal Images
Train Multimodel Speech
Train Multimodal Text
Accuracy Comparison Graph
Predict Emotion from Webcam
Predict Speech Emotion
Predict Text Emotion

```
VGG19 Speech Accuracy  :  98.25783972125436
VGG19 Speech Precision : 98.51326510333863
VGG19 Speech Recall    : 96.875
VGG19 Speech FScore    : 97.45325986694266

MobileNetV2 Speech Accuracy  :  97.21254355400697
MobileNetV2 Speech Precision : 97.52292334554917
MobileNetV2 Speech Recall    : 95.98577785696995
MobileNetV2 Speech FScore    : 96.5396541063503

ResNet50 Speech Accuracy  :  97.44773958236659
ResNet50 Speech Precision : 97.71562260508404
ResNet50 Speech Recall    : 96.71134910081054
ResNet50 Speech FScore    : 97.0606566071244

Xception Speech Accuracy  :  98.32869080779945
Xception Speech Precision : 98.39471205148726
Xception Speech Recall    : 97.66579609639257
Xception Speech FScore    : 97.94386960148019
```

Fig.4.6 performance of all algorithms on Speech dataset

In above screen can see performance of all algorithms on Speech dataset and now click on 'Train Multimodal Text' button to train all algorithms on TEXT data and get below output


Multimodal Emotion Recognition
Depression Detection from Text, Image & Speech using Deep Learning Algorithm
Upload Emotion Dataset
Preprocess Dataset
Train Multimodal Images
Train Multimodel Speech
Train Multimodal Text
Accuracy Comparison Graph
Predict Emotion from Webcam
Predict Speech Emotion
Predict Text Emotion

```
VGG19 Text Accuracy  :  97.7
VGG19 Text Precision : 97.68665047475382
VGG19 Text Recall    : 97.72006534691525
VGG19 Text FScore    : 97.69856160100063

MobileNetV2 Text Accuracy  :  98.6
MobileNetV2 Text Precision : 98.59329714496165
MobileNetV2 Text Recall    : 98.60497149080658
MobileNetV2 Text FScore    : 98.59873886497847

ResNet50 Text Accuracy  :  99.1
ResNet50 Text Precision : 99.10236233085446
ResNet50 Text Recall    : 99.09587417515536
ResNet50 Text FScore    : 99.09901883150751

Xception Text Accuracy  :  99.3
Xception Text Precision : 99.31040422369902
Xception Text Recall    : 99.28967262476776
Xception Text FScore    : 99.29914144827411
```

Fig.4.7 see performance of all algorithms on TEXT data

In above screen can see performance of all algorithms on TEXT data and then click on 'Accuracy Comparison Graph' button to get below graph

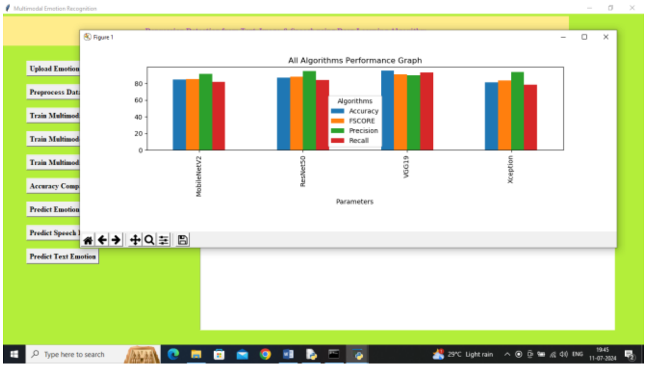


Fig.4.8 Graphical representation

In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms VGG19 got better performance for facial emotion. Now close above graph and then click on 'Predict Emotion from Webcam' button to get below output

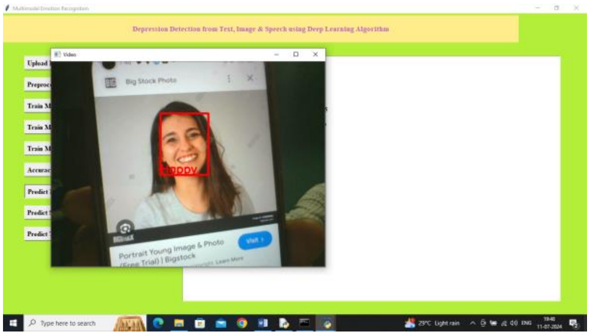

Depression Detection from Text, Image & Speech using Deep Learning Algorithm
Video
Big Stock Photo
happy
Portrait Young Image & Photo (Free Trial) | Bigstock
Visit

Fig.4.9 webcam started

In above screen webcam started and now show your faces to webcam to detect emotion and now click on 'Predict Speech Emotion' button to get below output

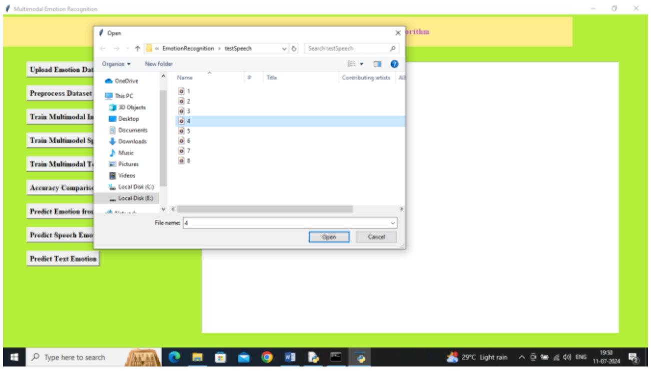
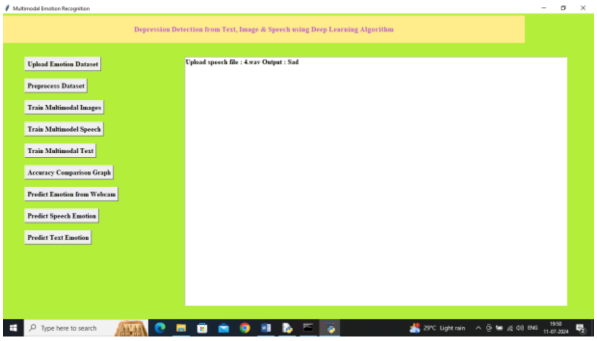


Fig.4.10. select and upload 'audio' file

In above screen select and upload 'audio' file and then click on 'Open' button to get below output



Upload speech file : 4.wav Output : Sad

Fig.4.11 text area output can see uploaded audio file

In above screen in text area output can see uploaded audio file emotion predicted as ‘Sad’ and similarly you can upload and test other audio files. Now click on ‘Predict Text Emotion’ button to upload ‘text data file’ and get below output

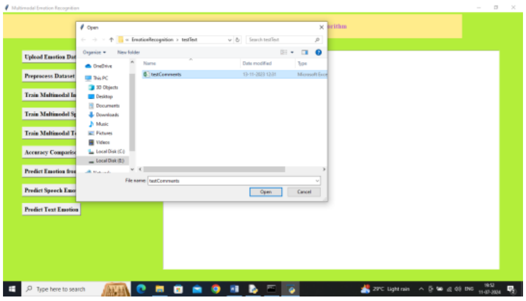
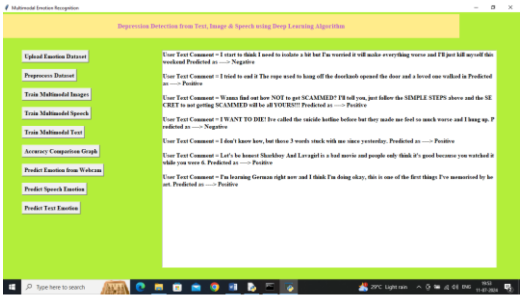


Fig.4.12 selecting and uploading ‘test comments’ file

In above screen selecting and uploading ‘test comments’ file and then click on ‘Open’ button to get below output



```
User Text Comment = I start to think I need to isolate a bit but I'm worried it will make everything worse and I'll just kill myself this
weekend Predicted as ----> Negative

User Text Comment = I tried to end it The rope used to hang off the doorknob opened the door and a loved one walked in Predicted
as ----> Positive

User Text Comment = Wanna find out how NOT to get SCAMMED? I'll tell you, just follow the SIMPLE STEPS above and the SE
CRET to not getting SCAMMED will be all YOURS!!! Predicted as ----> Positive

User Text Comment = I WANT TO DIE! Ive called the suicide hotline before but they made me feel so much worse and I hung up. P
redicted as ----> Negative

User Text Comment = I don't know how, but those 3 words stuck with me since yesterday. Predicted as ----> Positive

User Text Comment = Let's be honest Sharkboy And Lavagirl is a bad movie and people only think it's good because you watched it
while you were 6. Predicted as ----> Positive

User Text Comment = I'm learning German right now and I think I'm doing okay, this is one of the first things I've memorised by he
art. Predicted as ----> Positive
```

Fig.4.13 input test sentence from file

In above screen before arrow symbol --→ can see input test sentence from file and after arrow symbol can see predicted emotion as positive and negative.

So by using above application we have done emotion prediction using multiple data versions.

CONCLUSION

In this study on multimodal emotion recognition, we employed advanced convolutional neural networks—VGG19, MobileNetV2, ResNet, and Xception—to analyse and interpret human emotions from diverse data sources. Our results demonstrated high accuracy rates across these models, with Xception achieving the highest accuracy of 98.32%, followed closely by VGG19 at 98.25%, ResNet at 97.44%, and MobileNetV2 at 97.21%. These findings underscore the effectiveness of combining multiple advanced AI algorithms for emotion recognition tasks. The superior performance of the Xception model highlights its capability to integrate and process multimodal data efficiently.

REFERENCES

1. Sebe, Nicu, Ira Cohen, and Thomas S. Huang. "Multimodal emotion recognition." In *Handbook of pattern recognition and computer vision*, pp. 387-409. 2005.
2. Haq, Sanaul, and Philip JB Jackson. "Multimodal emotion recognition." In *Machine audition: principles, algorithms and systems*, pp. 398-423. IGI global, 2011.
3. Abdullah, Sharmeen M. Saleem Abdullah, Siddeeq Y. Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. "Multimodal emotion recognition using deep learning." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 73-79.
4. Lian, Hailun, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face." *Entropy* 25, no. 10 (2023): 1440.
5. Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal speech emotion recognition using audio and text." In *2018 IEEE spoken language technology workshop (SLT)*, pp. 112-118. IEEE, 2018.
6. Pagé Fortin, Mathieu, and Brahim Chaib-draa. "Multimodal multitask emotion recognition using images, texts and tags." In *Proceedings of the ACM Workshop on Crossmodal Learning and Application*, pp. 3-10. 2019.
7. Zhang, Xue, Ming-Jiang Wang, and Xing-Da Guo. "Multi-modal emotion recognition based on deep learning in speech, video and text." In *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, pp. 328-333. IEEE, 2020.
8. Koromilas, Panagiotis, and Theodoros Giannakopoulos. "Deep multimodal emotion recognition on human speech: A review." *Applied Sciences* 11, no. 17 (2021): 7962.
9. Kumar, Puneet, Sarthak Malik, and Balasubramanian Raman. "Interpretable multimodal emotion recognition using hybrid fusion of speech and image data." *Multimedia Tools and Applications* 83, no. 10 (2024): 28373-28394.

10. Siddiqui, Mohammad Faridul Haque, and Ahmad Y. Javaid. "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images." *Multimodal Technologies and Interaction* 4, no. 3 (2020): 46.
11. Mittal, Trisha, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, pp. 1359-1367. 2020.
12. Jia, Ning, Chunjun Zheng, and Wei Sun. "A multimodal emotion recognition model integrating speech, video and MoCAP." *Multimedia Tools and Applications* 81, no. 22 (2022): 32265-32286.
13. Caschera, Maria Chiara, Patrizia Grifoni, and Fernando Ferri. "Emotion classification from speech and text in videos using a multimodal approach." *Multimodal Technologies and Interaction* 6, no. 4 (2022): 28.
14. Bänziger, Tanja, Didier Grandjean, and Klaus R. Scherer. "Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT)." *Emotion* 9, no. 5 (2009): 691.
15. Hosseini, S., M. R. Yamaghani, and S. Poorzaker Arabani. "A review of the methods of recognition multimodal emotions in sound, image and text." *International Journal of Applied Operational Research-An Open Access Journal* 12, no. 1 (2024): 29-41.