## EMOTION RECOGNITION USING

## MULTI – MODAL DATA

## MSBA 326 - Machine Learning for Predictive analytics

## Golden Gate University, San Francisco

## Project Supervisor

## Dr. Bahman Zohuri

## Submitted By

## Hamsaveni Mani (ID:0608090)

## Hisham Kochirikunnu (ID: 0605578)

## Hrushikesh Jadhav (ID:0607849)

## Sai Teja Rekha (ID:0607857)

## Sumanth Ranga (ID: 0607850)

## Acknowledgment

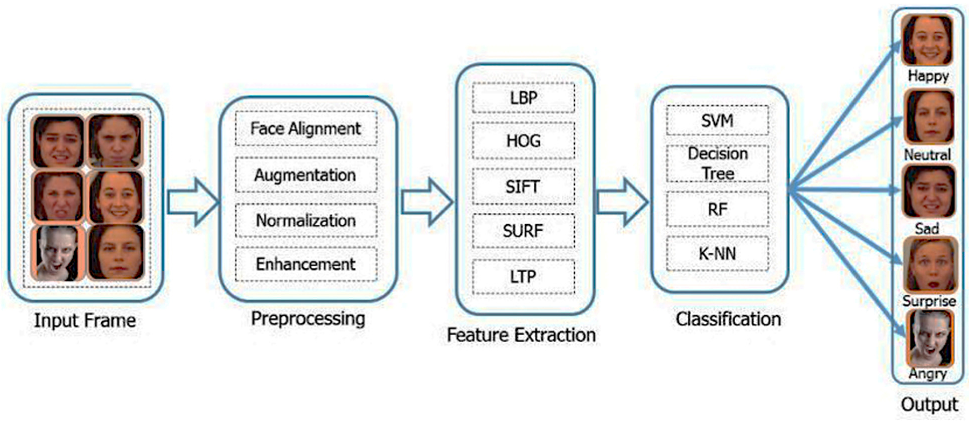
We would like to express our sincere thanks and special gratitude to our professor

Dr. Bahman Zohuri for leading us through this comprehensive research project “Emotion recognition using Multi-Modal Data” and for giving us his valuable guidance and insight throughout.

## Abstract

Facial expressions, as opposed to spoken words, visually transmit a great deal of information. Recognising facial expressions is important for human-machine interaction. Applications for the automatic facial expression recognition system are numerous and include, but are not limited to, comprehending human behaviour and creating artificial human expressions. It is still difficult for computers to recognise facial expressions with a high recognition rate.

Geometry and appearance-based methods are two widely used approaches. Pre-processing, face detection, feature extraction, and classification are the four steps that typically make up facial expression recognition.



We’ve used a variety of deep learning techniques for all three of the emotions used in this project which are classified separately as below:

Text: Go-emotions dataset

Audio/Speech:  RAVDESS dataset

Video/Image: RAVDESS and CK+ datasets

**TABLE OF CONTENT**

1. **Introduction 5**
   1. **The objective of the Study 6**
   2. **Limitation of the study 6**
   3. **Significance of the study 6**
2. **Literature Study 7**
3. **Understanding Emotion Recognition 8**
   1. **Importance of Emotion Recognition 8**
   2. **Modalities of Emotion Recognition 9**
4. **Building the Emotion Recognition Model 11**
   1. **Facial Emotion Recognition 11**
   2. **Speech Emotion Recognition 13**
   3. **Text Emotion Recognition 15**
5. **Research Methodologies Used 18**
   1. **Software used 18**
   2. **Details of Implementation 18**
   3. **Packages and Libraries used 18**
6. **Evaluating Model Accuracy 19**
   1. **Performance metrics 19**
7. **Conclusion 23**
8. **References 24**

**I. Introduction**

Emotion recognition technology has become a cornerstone in the realm of artificial intelligence, enabling machines to understand human emotions through various modalities such as facial expressions, speech patterns, and text analysis.

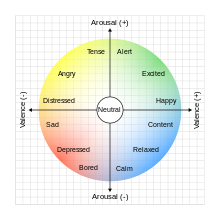
The branch of signal processing known as image processing deals with signals that have images as both their input and output. Face expression recognition is one of the most significant uses of image processing. Our facial expressions convey the emotions we are feeling. In interpersonal communication, facial expressions are crucial.

Facial expression is a scientific nonverbal gesture that conveys our feelings through our faces. This generation needs automatic facial expression recognition since it is essential to robotics and artificial intelligence. Applications such as videophone and teleconferencing, forensic software, human-computer interaction, automated surveillance, cosmetology, and personal identification and access control are some that are associated with this.

In this article, we will delve into the intriguing world of emotion recognition and explore the methodologies involved in integrating facial, speech, and text analysis using machine learning algorithms.

The main motive of this project is to develop this Facial Recognition System which can take this expression as input and recognize and classify it into seven different expression class such as:

For example: Calm, happy, Sad, angry, fearful, disgust, surprised



Since most studies have already been conducted in this way, our goal is to enhance the emotion recognition using Multi- Modal data while also increasing its accuracy when compared to other systems.

1. **The objective of the Study**

The main objective of the emotion recognition model is to identify a person's emotions. Text, Audio and Visual modes are used to capture emotion. Expression, words and facial gestures all help distinguish a people's feelings.

1. **Limitations of the Study:**

Some common limitations identified in the Emotion recognition model include:

**Integration and Accuracy**: Emotion recognition algorithms still require improvements in accuracy, and there is a need for a common affect representation model or mapping between models to integrate and compare results based on labels.

**Cultural and Semantic Challenges**: Emotion recognition systems face challenges related to cultural presumptions and the definition of emotions. Designing systems without imposing cultural presumptions about how emotions are defined is a complex task.

**Technical Challenges**: Technical challenges, such as the accuracy and disturbance robustness of emotion recognition algorithms, need to be addressed to enhance the effectiveness of these systems.

**Study Limitations**: Some studies acknowledge limitations such as a limited number of participants, arbitrarily chosen metrics and thresholds, and the need for additional case studies and experiments to validate the findings.

These limitations highlight the need for ongoing research and development to address the technical, cultural, and accuracy challenges associated with emotion recognition models.

1. **Significance of the Study:**

The significance of emotion recognition models is evident in various fields, as discussed in the provided sources. Emotion recognition can be realized through the analysis of facial expressions, speech, behaviour, or physiological signals. The integration of philosophy and machine learning theory has led to the development of facial expression emotion recognition models, which are crucial for understanding a person's mental state and facilitating interpersonal communication.

Despite criticisms and limitations, companies continue to invest in emotion recognition technology due to its potential applications in various domains, including human-computer interactions, healthcare, education, and the creative arts. However, the effectiveness of emotion recognition models remains a subject of debate, with challenges related to cultural presumptions, semantic definitions of emotions, and the accuracy and robustness of the algorithms.

While the technology has seen numerous applications, there is ongoing research to address the technical, organizational, and cost-related challenges associated with emotion recognition.

**II. Literature Study**

In the last few years, the deep learning computing paradigm has been deemed the Gold Standard in the machine learning community. Moreover, it has gradually become the most widely used computational approach in the field of machine learning, thus achieving outstanding results on several complex cognitive tasks, matching or even beating those provided by human performance. In particular, this paper outlines the importance of machine learning techniques and networks used for recognizing human emotions by 3 different models such as for speech, text and visual methods. It then presents convolutional neural networks (CNNs) which the most utilized deep learning network type and describes the development of CNNs architectures together with their main features, RAVDESS and CK+ datasets.

Previous works are focused on eliciting results from unimodal systems. Machines used to predict emotion by only facial expressions or only vocal sounds. After a while, multimodal systems that use more than one features to predict emotion has more effective and gives more accurate results. So that, the combination of features such as audio-visual expressions, EEG, body gestures have been used since. More than one intelligent machine and neural networks are used to implement the emotion recognition system. Multimodal recognition method has proven more effective than unimodal systems. Research has demonstrated that deep neural networks can effectively generate discriminative features that approximate the complex non-linear dependencies between features in the original set. These deep generative models have been applied to speech and language processing, as well as emotion recognition tasks showed that bidirectional Long Short Term Memory(BLSTM) network is more effective that conventional SVM approach.; In speech processing proposed and evaluated deep networks to learn audio-visual features from spoken letters. In emotion recognition, Brueckner et al found that the use of a Restricted Boltzmann Machine (RBM) prior to a two-layer neural network with fine-tuning could significantly improve classification accuracy in the Interspeech automatic likability classification challenge.

The work by Stuhlsatz et al. took a different approach for learning acoustic features in speech emotion recognition using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs). Emotion Recognition is an important area of work to improve the interaction between human and machine. Complexity of emotion makes the acquisition task more difficult. Quondam works are proposed to capture emotion through unimodal mechanism such as only facial expressions or only vocal input. More recently, inception to the idea of multimodal emotion recognition has increased the accuracy rate of the detection of the machine.

Moreover, deep learning technique with neural network extended the success ratio of machine in respect of emotion recognition. Recent works with deep learning technique has been performed with different kinds of input of human behavior such as audio-visual inputs, facial expressions, body gestures, EEG signal and related brainwaves. Still many aspects in this area to work on to improve and make a robust system will detect and classify emotions more accurately. In this paper, we tried to explore the relevant significant works, their techniques, and the effectiveness of the methods and the scope of the improvement of the results. visual emotion recognition process.

With all the above improvements and extensive research in CNNs, it is an extremely useful technique for image analysis, pattern recognition, or extracting features applications (Kulkarni et al., 2009) Just after the release of FER2013, a huge face recognition data, during ICML in 2013, it has become a benchmark when comparing the performance of the model in emotions recognition. Several CNN variants had achieved excellent results, with recognition rates varying between 65 through 72.7 %, for instance, training three separate CNN models and merged these to improve performance. The single best efficiency they've gotten is 62.44%, using a perceptual convolutional layer inside an edge deep learning model achieved an efficiency of 70.02 %. Thus, it's been shown that integrating multiple models improves overall efficiency.

Similarly, in emotion recognition model, Emotions may be divided into various categories. Classifications like happiness, sadness, disgust, anger, and fear, and also. Text, speech, and visual are all types of information. The combining of such data gives the most precise efficiency. That form of fusing could be advantageous.

**III. Understanding Emotion Recognition**

Emotion recognition, also known as affective computing, involves the development of algorithms that enable machines to identify and interpret human emotions. This technology is instrumental in human-computer interaction, mental health diagnostics, and customer sentiment analysis.

1. **Importance of Emotion Recognition**

Emotion recognition plays a pivotal role in enhancing user experiences, making applications more intuitive, and fostering empathetic artificial intelligence systems. The importance of emotion recognition can be summarized as follows in various fields:

**Adaptive Social Behaviour**: Recognition of emotion from facial and body expressions is crucial to adaptive social behaviour. Emotion recognition guides response and action toward potential friendly or threatening others and is paramount to successful communication between individuals

**Evolutionary Mechanism**: Emotion recognition plays a vital role in our evolutionary mechanism by keeping us aware of the situation and the safe, smart, and adequate ways to react to it

**Effective Communication and Healthy Relationships**: Effective communication and healthy relationships with others are a two-way street of mutual respect and understanding. Emotion recognition helps in interpreting and reacting to the hints given out by others, voluntarily or not

**Various Applications in Different Fields**: Emotion recognition has found applications in human-computer interactions (HCI), medical health, internet education, security monitoring, intelligent cockpits, psychoanalysis, and the entertainment industry

**Detection of Lies and Mental States**: Emotion recognition can be used to detect lies (authenticity test) and detect the drowsiness and mental state of drivers to improve driving safety

**Assistance for People with Special Needs**: Emotion recognition can be applied to recognize the emotions of the elderly, infants, and those with special diseases who cannot clearly express their emotions

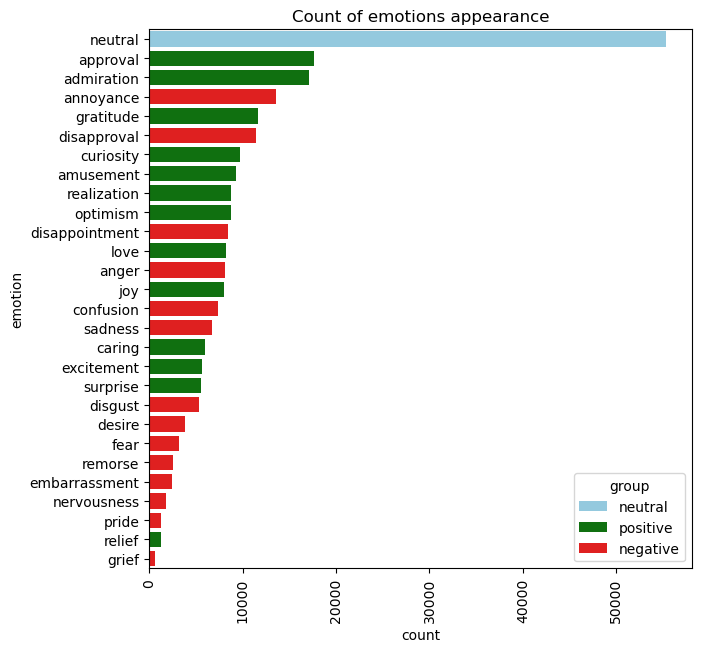
**Improvement in Medical Efficiency and Experience**: In the field of medical and healthcare, emotion recognition can be used to detect a patient's psychological state or adjuvant treatment, and improve medical efficiency and medical experience

Overall, emotion recognition is a significant area of research with numerous applications in various fields. It plays a crucial role in human interactions, communication, and understanding emotions, which can be beneficial for improving mental health, driving safety, and enhancing medical experiences.

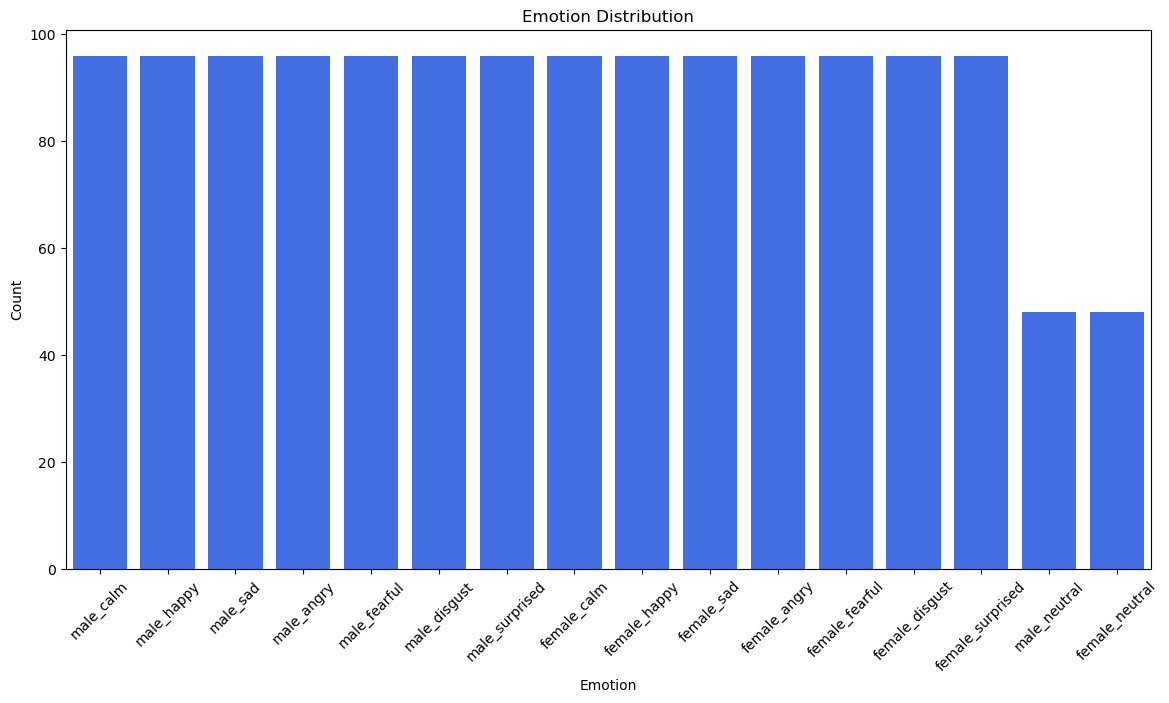
**b) Modalities of Emotion Recognition**

The different modalities used for emotion recognition in this project involves Text, Audio/speech and Visual/facial which are explained below:

**Text Analysis**: Parsing textual data to understand the sentiment and emotional context.



**Audio/ Speech Analysis**: Analysing speech patterns, tone, and intonation to identify emotional states.

****

**Visual/ Facial Analysis**: Utilizing facial expressions to decipher emotions such as happiness, sadness, anger, calm and surprise.

First in text modality, sentiment analysis technique is used which is nothing but guessing the mood of the writer sometimes when we read a text. It uses dictionaries of words that have feelings attached to them. Even the way we write—like using exclamation marks or emoticons—can show emotions. We teach our system to notice these little clues.

Words have power and context. By understanding how words are used together, we get a deeper sense of the emotions they convey. That's where word embeddings come in, mapping words in a way that captures their emotional context.

Next in audio Modality which uses prosodic features like pitch, energy and speaking rate to recognize emotions. Our voices reveal a lot about how we feel. Changes in pitch, how loud we speak, and how fast we talk can all show different emotions. Our system listens for these changes. The quality of our voice—whether it's smooth or rough—can also tell a story. We analyse these spectral features to understand emotions better.

There are tools already made, like wav2vec 2.0, that are great at understanding sounds. We use these to give our system a head starts in recognizing emotions from audio.

Finally in Visual Modality: using this model a lot can be said with just a smile or a frown. Our faces show emotions through different muscle movements. We teach our system to recognize these facial action units. Not just the movements, but also the key points on our face, like the corners of our eyes or mouth, give away our feelings. These facial landmarks are crucial for our system."

Finally, we use powerful image processing tools, like ResNet , which are types of deep learning models. They help our system understand the complex patterns in face images and videos.

**IV. Building the Emotion Recognition Model**

Building an emotion recognition model involves several crucial steps, each tailored to specific modalities. In this paper we’ll see about the Facial/ visual recognition model, Speech/ sound recognition model and Text recognition model which are briefed below

1. **Facial Emotion Recognition model:**

Data Collection: Gathering diverse facial expression datasets for training the model.

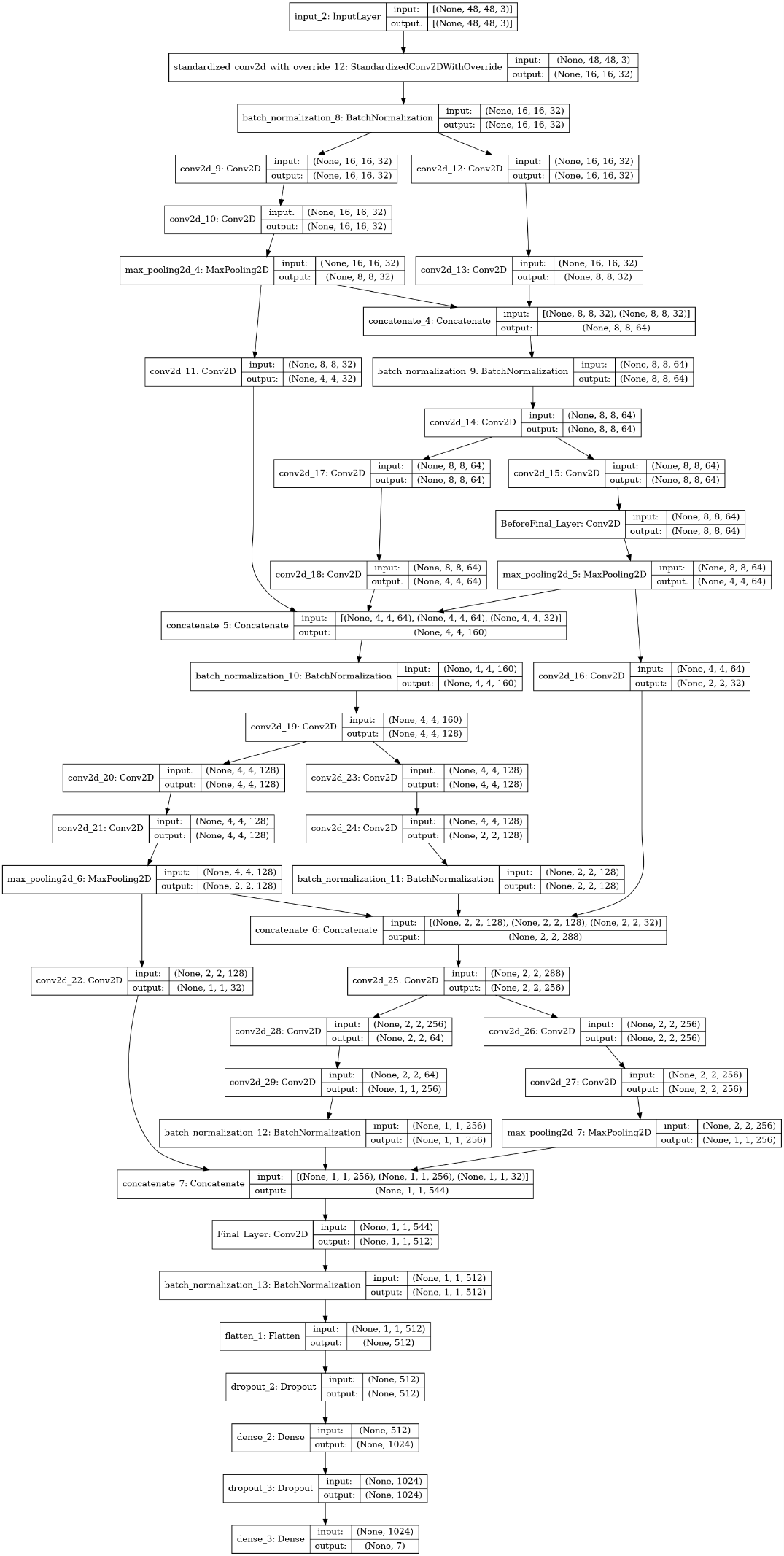
Pre-processing: Normalizing images, detecting facial landmarks, and extracting relevant features.

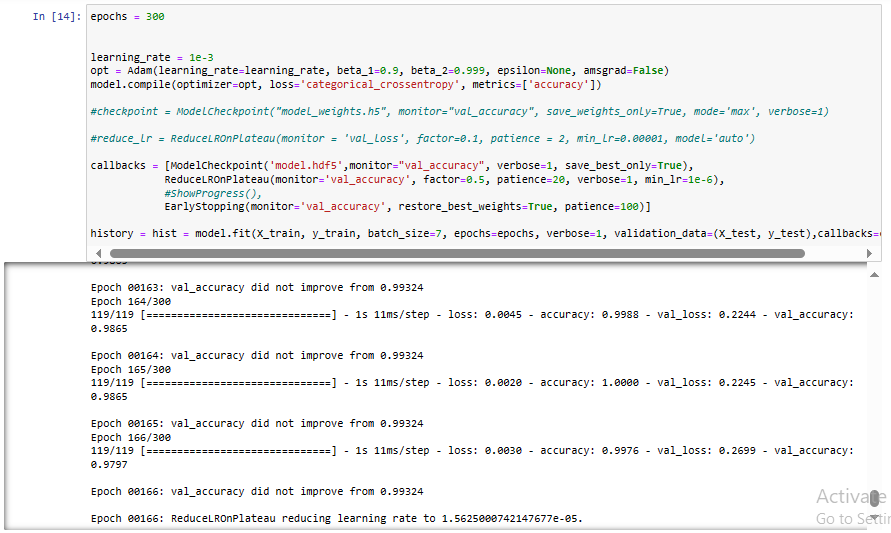
Model Selection: Choosing appropriate deep learning architectures such as Convolutional Neural Networks (CNNs) for facial emotion recognition.







****

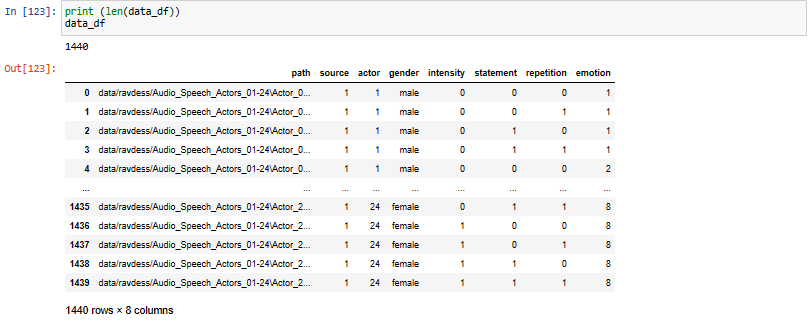


1. **Speech Emotion Recognition model:**

Data Acquisition: Collecting audio datasets encompassing various emotional states and contexts.

Feature Extraction: Extracting features like Mel-frequency cepstral coefficients (MFCCs) from audio signals.

Model Training: Employing Convocational neural networks (CNNs) for speech emotion recognition.

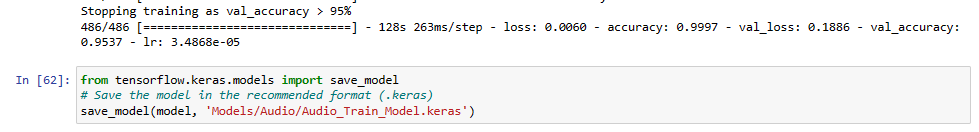












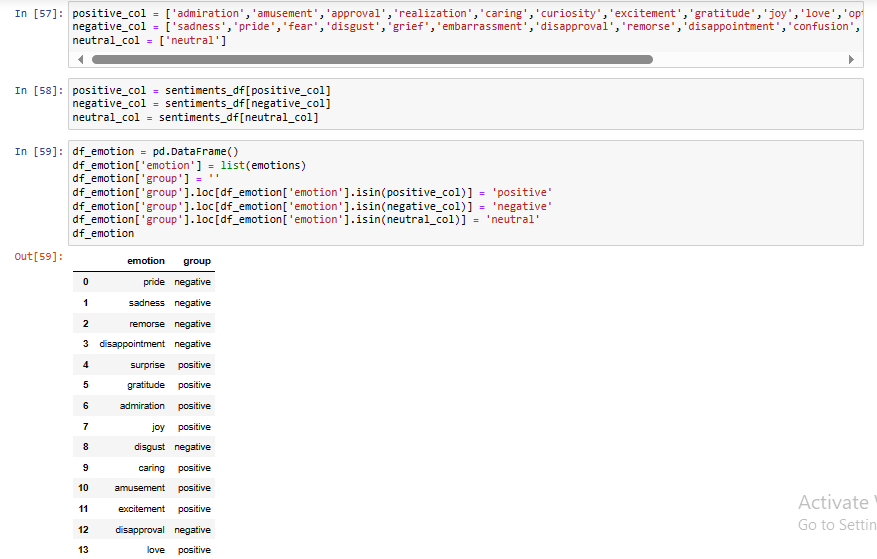
**C) Text Emotion Recognition model:**

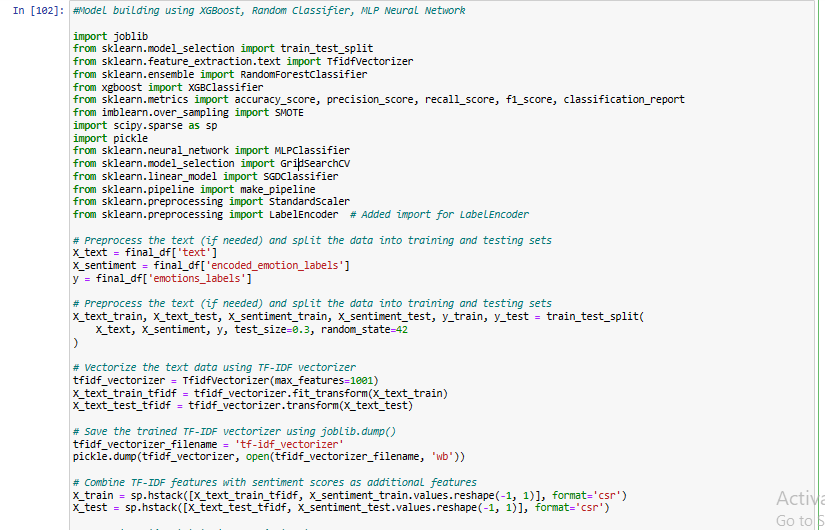
Dataset Preparation: Compiling textual datasets annotated with emotional labels.

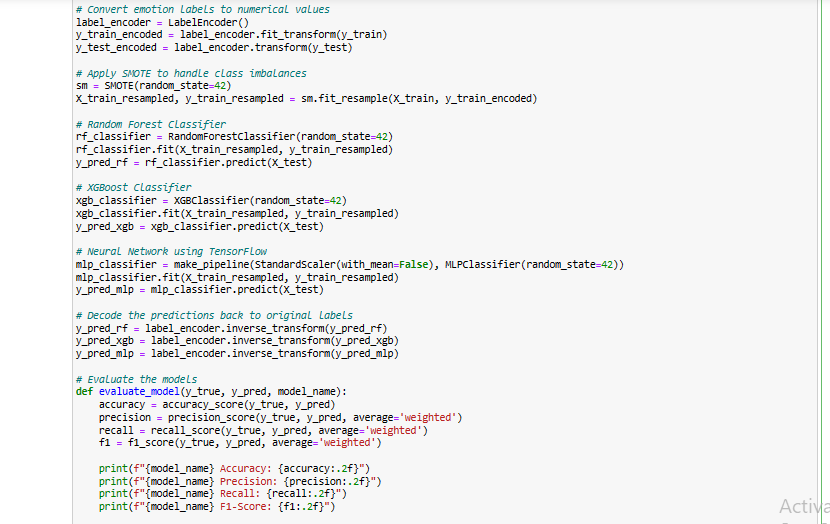
Text Vectorization: Converting text into numerical vectors using techniques like Word2Vec or TF-IDF.

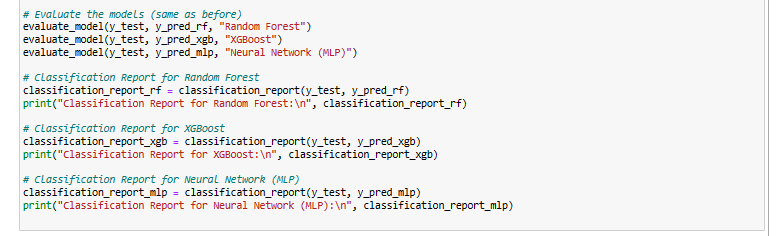
Model Implementation: Employing natural language processing (NLP) models like LSTM or Transformer for text emotion recognition.











**V. Research Methodology**

**a) Software Used:**

We have used Anaconda for Python 3 because the project is written in Python.

1. Anaconda: It was a free software Python and R computer language distribution for data science and machine learning techniques (big data processing, data modeling, computer science) that seeks to make package development and delivery simpler. Conda, a package management solution, keeps a record of package versions. About six million people just use the Anaconda service that contains over 250 popular Windows, Linux, and macOS machine learning bundles

2. Google collab: The Colab, sometimes owned by Google Colaboratory, is indeed a free machine learning platform sponsored by Google. It gives its users access to free CPU & GPU capabilities. It's perfect for people & small companies involved in Computer Vision who have not had recourse to a GPU. COLAB operates using Jupyter Notebook data and needs no setup. The following are some of the features: 1. Source code editor with dynamic typing & insight 2. Several Python terminals are available (including IPython) 12 12 3. A graphical user interface (GUI) for discovering & changing constants The following plugins are accessible: 1. Palin to Coding Profiler för Code Coverage Evaluation 2. Conda Package is a programming framework for both the Conda system.

**b) Details of Implementation:**

1.Dataset: The database used to train the model came from either a prior, RAVDESS and CK+. The photographs are grayscale images of faces with a resolution of 48x48 pixels. In the dataset, there are the following emotion labels: Calm, happy, Sad, angry, fearful, disgust, surprised

**c) Packages & Library**:

1. OpenCV: OpenCV is an open-source computer vision and machine learning library. OpenCV was established to offer a common infrastructure for computer vision applications and to accelerate the introduction of machine perception into consumer products. Since OpenSSL is a Cygwin software, it's indeed simple for enterprises to use. And over 2500 tailored equations are included in the collection, which includes a comprehensive set both of traditional or slicing image processing and artificial intelligence algorithms. This architecture might be used to recognize and identify photographs, categorize human actions in films, devices employ motions, track movement, extract 3-dimensional items, construct 3d images of particles photographs, connect pictures just to create a high-resolution representation of a whole program, and so much more. Find similar images in a database, remove dark hair from flashing images, track eye movements, identify landscape or position markers to add this with augmented reality, and so forth. OpenCV does have a user community of over 47 thousand people as well as a total number of likes of over 14 million. Companies, study organizations, and federal agencies all make extensive use of the library. It's available for Windows, Linux, Android, & Mac OS and also has C++, Python, Java, & MATLAB frameworks. When MMX & SSE instructions are accessible, OpenCV mainly focuses on true computer vision. OpenCV was created in C and C++, and it features a functionalized design that neatly integrates with the STL container.

2. NumPY: NumPy is the acronym for "Numerical Python" or "Numeric Python." It was an extension module for Python that offers precompiled functions for statistical and mathematical operations. NumPy frequently brings robust database systems to the Python programming language, enabling quicker inter-array computation. Large matrices and arrays are also the implementation's goal. The unit provides a vast array of sophisticated computations to operate on such matrices and sequences as well. The most important Python module for computer science is this one.

3. Tensor Flow: Google developed and published TensorFlow, a Python library for fast numerical computation. It was a base library that can be used to construct Deep Neural networks direct or via wrapper dependencies placed on top of TensorFlow to make things easier.

4. Panda: The Pandas library is a powerful and widely used open-source data analysis and manipulation tool for the Python programming language. It offers a range of features and functionalities that make it easier to work with and analyze data. Some key features and functionalities of the Pandas library include: Data Structures, Data Import & Export, Data cleaning & Manipulation, Data Analysis, Integration with other Libraries and community support. Overall, the Pandas library is a valuable tool for anyone working with data in Python, offering a wide range of features and functionalities that make it easier to perform analytical tasks and derive insights from data.

**VI. Evaluating Model Accuracy**

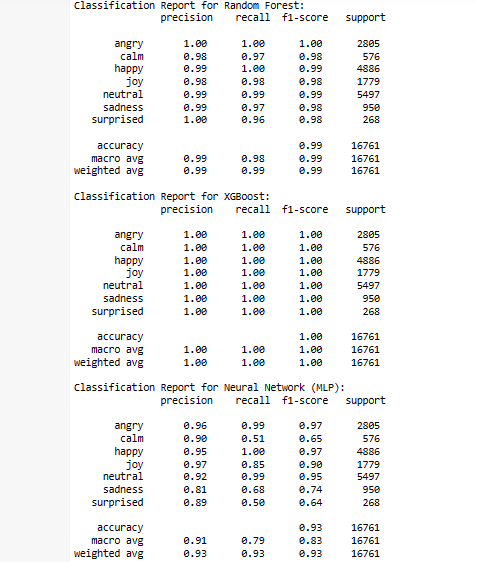
One of the challenges for evaluating emotion recognition systems is the choice of appropriate performance metrics and benchmarks. Depending on the task and the data, different metrics may be used to measure the accuracy, precision, recall, or f1-score of the system. For example, accuracy may be suitable for discrete emotion classification, but not for continuous emotion regression. Moreover, different benchmarks may have different levels of difficulty, quality, and diversity of the data.

For example, some benchmarks may use controlled laboratory settings, while others may use naturalistic or in-the-wild scenarios. Therefore, machine learning researchers may need to compare and contrast their results with multiple metrics and benchmarks, and report the limitations and assumptions of their methods. However, the performance metrics for this model is given below:

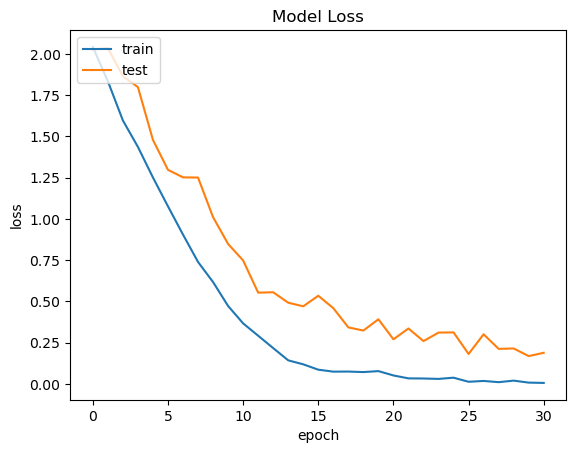
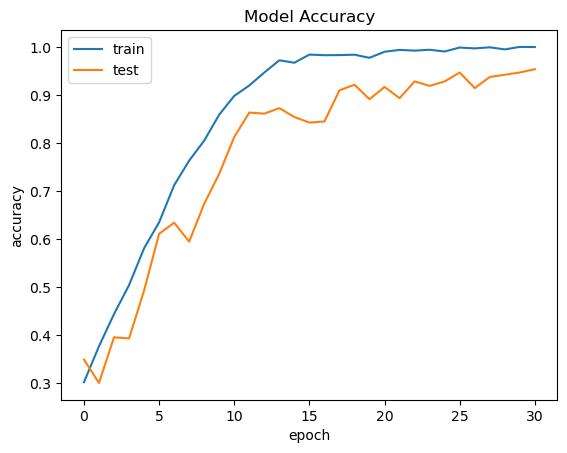
1. **Performance Metrics**

* Accuracy: Measuring the overall correctness of emotion predictions across modalities.
* Confusion Matrix: Analyzing true positive, true negative, false positive, and false negative predictions.
* F1 Score: Balancing precision and recall for a holistic evaluation.

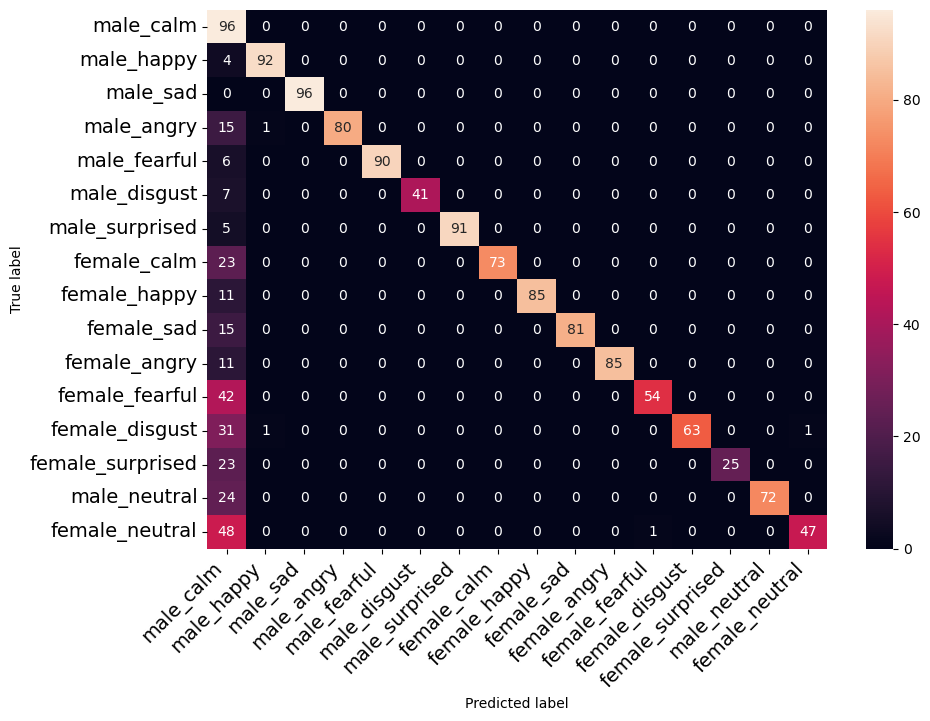
**Text Metrics:**



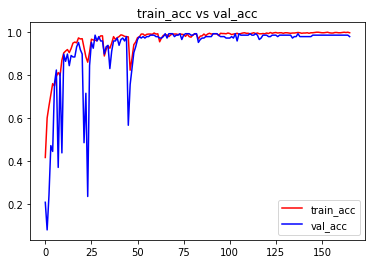
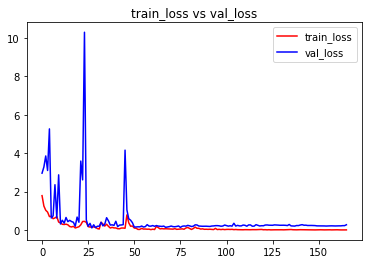
**Speech Metrics:**

****



****

**Facial Metrics:**





**VII. Conclusion**

In conclusion, the integration of facial, speech, and text analysis in emotion recognition projects is a promising avenue in artificial intelligence research. By employing advanced machine learning techniques and diverse datasets, developers can create empathetic and emotionally intelligent applications.

Specifically in this Paper we’ve discussed about the work done on emotion recognition and the models used for achieving it. We have proposed a glimpse of a probable solution and method towards recognizing the emotion using 3 models which are text, speech and visuals.

In the carried-out experiments, for 7 emotional states such as Calm, happy, Sad, angry, fearful, disgust, surprised, we achieved a very good accuracy of emotions using image which is 99.3%, similarly using text we got an excellent accuracy rate of 100% and for audio we got a accuracy rate of – 81.32%. Certainly, the accuracy would be impacted by various factors which can impact the accuracy.

**VIII. References**

A literature survey on Facial Expression Recognition using Global FeaturesbyVaibhavkumar J. Mistry and Mahesh M. Goyani, International Journal of Engineering and Advanced Technology (IJEAT), April 2013 [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.645.5162&rep=rep 1&typ e=pdf]

Emotion Recognition Based on Multimodal Information [Zhihong Zeng](https://link.springer.com/chapter/10.1007/978-1-84800-306-4_14#auth-Zhihong-Zeng), [Maja Pantic](https://link.springer.com/chapter/10.1007/978-1-84800-306-4_14#auth-Maja-Pantic) & [Thomas S. Huang](https://link.springer.com/chapter/10.1007/978-1-84800-306-4_14#auth-Thomas_S_-Huang) [https://link.springer.com/chapter/10.1007/978-1-84800-306-4\_14]

A review on deep convolutional neural networks, [Neena Aloysius](https://ieeexplore.ieee.org/author/37086268492); [M. Geetha](https://ieeexplore.ieee.org/author/37949266500) [https://ieeexplore.ieee.org/abstract/document/8286426]

P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, Introduction to EEG- and Speech-Based Emotion Recognition. 2016. doi: 10.1016/C2015-0-01959-1.

Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, “MEC 2017: Multimodal Emotion Recognition Challenge,” 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia 2018. 2018. doi: 10.1109/ACIIAsia.2018.8470342.

S. Khan, “Systematic Literature Review for Facial Expression Recognition,” 2016.

S. Saganowski et al., “Emotion Recognition Using Wearables: A Systematic Literature Review-Work-in-progress,” in 2020 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2020, IEEE, Mar. 2020, pp. 1–6. doi: 10.1109/PerComWorkshops48775.2020.9156096.

C. Suman, R. Chaudhari, S. Saha, S. Kumar, and P. Bhattacharyya, “Investigations in Emotion Aware Multimodal Gender Prediction Systems From Social Media Data,” IEEE Trans Comput Soc Syst, pp. 1–10, 2022, doi: 10.1109/tcss.2022.3158605.

F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, “Recognizing emotion from speech based on age and gender using hierarchical models,” Procedia Comput Sci, vol. 151, pp. 37–44, 2019, doi: 10.1016/j.procs.2019.04.009.

Q. Su, F. Chen, H. Li, N. Yan, and L. Wang, “Multimodal emotion perception in children with autism spectrum disorder by eye tracking study,” 2018 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2018 - Proceedings, pp. 382–387, 2019, doi: 10.1109/IECBES.2018.8626642.

M. Rocha et al., “Towards Enhancing the Multimodal Interaction of a Social Robot to Assist Children with Autism in Emotion Regulation,” … on Pervasive Computing …, pp. 398–415, 2022, doi: 10.1007/978-3-030-99194-4\_25.