iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: XII Month of publication: December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57652

www.ijraset.com

Call: ✆08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Emotion Recognition through Text, Speech and Image

Dr. K. Shirisha Reddy[1], Shreejit Cheela[2], Vignya Durvasula[3]
[1]Associate Professor, HoD, [2]Student/Research Scholar, [3]Student/Research Scholar, Department of Computer Science and Engineering [Artificial Intelligence & Machine Learning], Vignana Bharathi Institute of Technology, Aushapur, Ghatkesar, Hyderabad.

Abstract: Emotion recognition plays a vital role in various applications, such as human-computer interaction, affective computing, and psychological assessment. This comparative study investigates the effectiveness of emotion recognition through text, image, and speech modalities. Our research aims to analyze and compare state-of-art approaches in each modality and identify their strengths and limitations. Conducting an exhaustive literature review enabled us to understand existing methodologies, datasets, and evaluation metrics. The research methodology and implementation include data collection, preprocessing, feature extraction, and the application of machine learning and deep learning models. The results provide insights into the performance of different modalities, paving the way for advancements in emotion recognition research.

Keywords: emotion recognition, text, image, speech, comparative study, performance evaluation.

I. INTRODUCTION

Emotions are crucial in human communication and perception, influencing our behavior and decision-making. As a result, emotion recognition has garnered significant interest in recent years. It involves identifying and classifying human emotions based on various modalities including text, image, and speech. Emotion recognition systems find applications in many fields such as human-computer interaction, social robotics, and mental health assessment.

A. Background of the Study

Emotion recognition has applications in various fields such as healthcare, education, and entertainment. Emotion recognition can be done through various methods such as facial expression recognition, physiological signals recognition, speech signals variation, and text semantics on standard databases such as JAFFE, CK+, Berlin Emotional Database, SAVEE, etc. as well as self-generated databases. Previous studies have shown that emotion recognition through text is more challenging than through speech or image. However, recent studies have shown that deep learning approaches can be used to improve the accuracy of Emotion Recognition.

B. Problem Statement

While individual studies have focused on emotion recognition through text, image, or speech, there is a lack of thorough comparative analyses. A thorough investigation is necessary to understand the strengths and limitations of each modality, identify the most effective techniques, and explore potential synergies between them.

C. Research Objectives

The primary objective of this study is to compare and evaluate the effectiveness of emotion recognition through text, image, and speech. Specific objectives include:

1) Analyzing state-of-the-art approaches in each modality.
2) Investigating the performance of machine learning and deep learning models.
3) Identifying common challenges and limitations in emotion recognition across modalities.
4) Exploring the potential complementarity of text, image, and speech-based approaches.

D. Research Questions

To achieve the research objectives, the study seeks to address the following research questions:

1) What are the current state-of-the-art approaches in emotion recognition through text, image, and speech?
2) How do different machine learning and deep learning models perform in each modality?

3) What are the common challenges and limitations in emotion recognition across modalities?
4) Are there potential synergies or complementarities between text, image, and speech-based approaches?

E. Significance of the Study

This research study is significant as it provides insights into the current state of emotion recognition through text, image, and speech. The findings will aid researchers and practitioners in selecting appropriate techniques for emotion recognition tasks and developing more accurate and robust emotion recognition systems. The study's outcomes contribute to the broader field of affective computing and advance the understanding of human-computer interaction.

II. LITERATURE REVIEW

A section providing insights into emotion recognition research progress and where the world stands as of now. It is an in-depth summary of the research done so far.

A. Previous Studies on Emotion Recognition Through Text, Image, and Speech

Detailed research has been conducted in emotion recognition through text, image, and speech modalities, yet challenges persist in achieving accurate recognition across various domains.

The authors of [1] explored various deep-learning algorithms to study the trends in text-based emotion recognition and found that bidirectional LSTMs outperform simple LSTMs. The model was found to do well when combined with bidirectional processing, dropout regularization, and weighted loss functions to manage the imbalances in the datasets. The idea of using Bi-LSTMs in our Text Emotion Recognition model has been inspired by this paper.

The authors of [2] worked on an end-to-end multimodal system that was operated on REmote COLlaborative and Affective (RECOLA) database to recognize spontaneous emotions using deep neural networks on raw speech and visual data. The proposed model showcased superior performance when compared to unimodal models, affirming the value of combining speech and visual data for emotion recognition.

In [3] the authors delved into various deep learning-based architectures including LSTM, CNNs, multi-layered Perceptron, and techniques such as Adam, Attention-based RNN encoders to operate on IEMOCAP dataset for speech, text, and facial data. Their multimodal model provided an accuracy of 71.04%.

The authors of [4] compared two multimodal emotion recognition models – deep canonical correlation analysis (DCCA) and bimodal deep autoencoder (BDAE) across five different emotion recognition datasets- SEED-IV, SEED-V, DEAP, MAHNOB-HCL, and AMIGOS. Wei Liu and their co-authors observed that DCCA achieves an average accuracy of 80.5% while BDAE achieved 77.3%. These results were also compared with a traditional approach which was 73.8% accurate. DCCA shows better robustness to noise and missing data than BDAE and traditional approaches.

The paper [5] proposes an approach to improve emotion recognition by using an attention mechanism and sequence model. The proposed approach is operated on the two modalities. The results show that the proposed approach achieves state-of-the-art performance on the IEMOCAP dataset. The proposed approach with the Oracle text achieves the best results in the dataset, showing that further improvement can be achieved by more accurate speech recognition.

The model proposed in [6] utilizes 3D-Convolutional Neural Networks for feature extraction which produces 3D feature maps using 3D filters. It also makes use of transfer learning and data augmentation to overcome the lack of data. The authors have also leveraged ensemble learning techniques such as bagging and score fusion to get the final fusion predictions.

The authors of [7] proposed a model called M3ER (Multiplicative Multimodal Emotion Recognition) that takes multiple input modalities such as face, text, and speech and is robust to sensor noise, produces proxy features. Furthermore, the authors claim that the proposed model outperforms the state-of-art algorithms such as LSTM, and DeepSpectrum on the two benchmark datasets: IEMOCAP, CMU-MOSEI in terms of F1-score and MA score.

As the title of [8] suggests, the proposed model utilizes cross attention to infer N-gram features extracted by two parallel branches – one using Bi-RNNs with DCNNs while the other uses DCNNs alone. The proposed system is claimed to outperform all the SOTA methods on IEMOCAP dataset. It achieved a weighted accuracy (WA) of 79.22% and unweighted accuracy of 80.51% using Fusion III.

Explained in [9] is a new approach called Progressive Modality Reinforcement (PMR) which includes a dynamic filter that selects features from each modality, and a message hub that facilitates modal interaction. The proposed approach outperforms the existing state-of-the-art models with an accuracy of 51.8%.

The authors of [10] conducted extensive research on emotion recognition and sentiment analysis using Convolutional Multiple Kernel Learning (MKL). For emotion recognition, they tested their model on the USC IEMOCAP dataset. Their proposed model was tested against visual and textual modalities and was observed to achieve significantly better results than SOTA methods for sad and neutral emotions with a confidence level of 95%.

In [11] the authors dived into emotion recognition from speech using various models of deep learning such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), Autoencoders, Deep Neural Networks (DNNs) and hybrid models. The performance of each model is compared with that of traditional machine learning techniques such as Gaussian Mixtures Models (GMMs) and Support Vector Machines (SVMs) on popular datasets like the IEMOCAP dataset, German-Aibo emotion and Smart-Kom dataset, RAVDESS dataset.

The model proposed in [12] leverages deep learning to recognize emotions from big data. The big data comprises speech and video. The speech signal is fed as an image to the Convolutional Neural Network (CNN). The frames of a video segment are also fed to another CNN and fused using two extreme learning machines (ELMs). The proposed model achieved a maximum accuracy of 99.9% with the help of an ELM fusion strategy with Big Data. The model was also tested on eNTERFACE database whose result was 86.4% with fusion.

The authors of [13] have conducted a detailed review of advancements of existing models in the field of facial emotion recognition (FER) using deep learning. The previous/existing models were trained and tested on several databases like MultiPie, MMI, GEMEP FERA, FER2013, JAFFE, RaFD, and many databases. The research conducted by the authors shows that high precision in FER can be achieved by applying CNN networks. A handful of new methods were proposed previously that combine CNN, RNN, and LSTM networks.

In [14] the authors came up with a new deep neural network model for Automated FER consisting of two convolutional layers, and four Inception layers. It is a single-component architecture and takes facial images as input and classifies them into six basic expressions or neutral. This proposed model was claimed to achieve superior results as compared to several SOTA methods.

The authors of [15] have leveraged the power of deep neural networks (DNNs) and Extreme Learning machines (ELMs) to learn and classify emotional information from low-level features. The DNN is built with three hidden layers and the number of hidden units in ELM is set to 120. The proposed model is set to achieve a 20% accuracy improvement compared to state-of-the-art approaches.

The authors of [16] proposed a neural network based on ResNet18 (SResNet18) to recognize facial expressions. The proposed approach was applied to the FER2013 and CK+ databases and compared with other state-of-the-art methods. The results showed that the proposed algorithm not only improves the accuracy but also reduces the size of the model, which is competitive with existing methods in terms of the size of the model parameters and recognition accuracy.

III. METHODOLOGY

A. Research Design

In this research, our primary objective was to conduct a complete study of existing emotion recognition models while also developing a multi-input model capable of accommodating three distinct types of inputs: text, image, and speech. With a foundational approach, the initial steps involved are data acquisition and preprocessing. Gratefully acknowledging the contributions of previous researchers, this study involved collecting multiple datasets from various reputable sources, including papers, repositories, and organizations like Hugging Face, Inc., to ensure data reliability and relevance. We cordially wish to ensure that all sources utilized in this study will be cited appropriately to maintain academic integrity.

The data acquisition process was followed by a meticulous data preprocessing and cleansing phase. Vital enhancements were incorporated into the datasets to ensure data quality and consistency. The details of the data preprocessing techniques employed will be elaborated in subsequent sections.

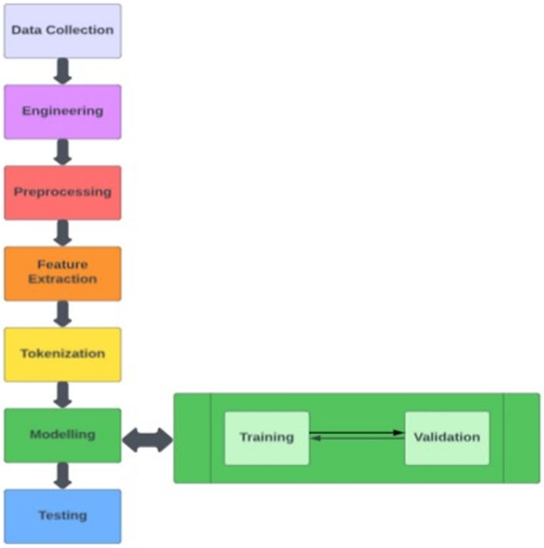
In the modeling phase, we designed and implemented individual machine learning models for each data modality. Specifically, a Bi-directional Long Short-Term Memory (Bi-LSTM) model was employed for text data, a classical machine learning model and a Simple ANN for audio data, finally a **Convolutional Neural Network (CNN)** for image data. Each model was built from scratch, tailored to leverage the unique characteristics of its respective data type.

Although we did not end up doing this, one of the key features of our research is to incorporate a multi-input model to recognize the input type and seamlessly direct it to the appropriate sub-model. For instance, when a text input is received, the text processing pipeline involving the Bi-LSTM model is invoked. Similarly, for audio and image inputs, the relevant processing pipelines with their respective models are triggered. This multi-functionality enables our model to handle mixed inputs and make accurate emotion predictions across different modalities.

While the presence of multiple pipelines may lead to some delay in processing, we believe the benefits of the unified model's versatility and the ability to handle varying data more than compensate for this trade-off. Our research design assimilates careful consideration of both efficiency and performance to strike a balance between accuracy and practicality.

Through this research, we aim to contribute to the field of emotion recognition and make this a one-stop destination for all and any kind of related information. By providing a complete description about the existing models and presenting a novel multi-input model that showcases the potential of integrating many and various modalities for more robust emotion analysis. Our work serves as a stepping-stone towards developing emotion recognition systems that can better understand human expressions and pave the way for more sophisticated applications in the realm of human-computer interaction and affective computing.

The process of data acquisition and model building is carried out as shown below.



B. Dataset

The significance of data cannot be overstated when it comes to the development of Machine Learning (ML) and Deep Learning (DL) models. Data serves as the bedrock upon which algorithms learn, generalize, and make informed decisions, propelling breakthroughs in various domains. Our research, focusing on Emotion Recognition, heavily relies on the quality, quantity, and diversity of the data used for training. By harnessing a comprehensive dataset that encompasses text, image, and speech data, we have successfully developed a robust Emotion Recognition model with exceptional performance in real-world applications. This section presents a detailed account of our data acquisition and processing steps.

The emotion classes selected for the three modules embrace a wide range of emotional expressions: happy, sad, angry, disgust, serene, and startle. The class "serene" serves as an umbrella term, encapsulating the emotions of both neutrality and calmness. Conversely, the "startle" class captures the emotional states of fear or surprise, providing a full-fledged representation of the conflicting emotions we aim to recognize and classify in our research.

1) Data Acquisition

As Tim O’Reilly wisely stated, "Who has the data, has the power" and with such power comes the great responsibility of careful data acquisition. The data acquisition process played a crucial yet challenging role in our research. While multiple approaches and metrics exist for gathering data, we adopted a meticulous approach, leveraging datasets from previous research, Kaggle, and online libraries. This approach, though tiresome, ensured the reliability and efficiency of our data acquisition process.

Metrics

a) Data Source and Reliability: Selecting a reliable data source was paramount to our data gathering process. We sought data from reputable sources, including well-established emotion-related databases, publicly available emotion recognition datasets, and validated emotion expression datasets. By relying on reputable sources, we ensured the accuracy, authenticity, and relevance of the data to our research objectives.

b) *Data Size:* The size of the dataset plays a vital role in the performance and generalization ability of machine learning models. We considered the complexity of the emotion recognition task and the available computational resources. To build an extensive dataset, we gathered a substantial volume of text, image, and speech data, incorporating various emotional expressions. This vast dataset, comprising millions of samples, empowered our model to learn from distinct emotional patterns.

c) *Data Balance:* A balanced dataset helps prevent bias in the model and ensures that each emotion is adequately learned and recognized. However, achieving perfect data balance across all three modules was challenging due to the inherent variability in emotional expressions in real-world data sources. To mitigate the impact of class imbalance during model training, we employed data augmentation techniques, which artificially balanced the dataset to some extent.

d) *Data Diversity:* Data diversity was a crucial consideration during our data gathering process. We endeavored to include data from wide-ranging demographics, cultures, and regions, ensuring that our Emotion Recognition model could generalize effectively across different populations. By incorporating data with unlike emotional expressions, we aimed to enhance the model's ability to recognize emotions in various contexts and across various user groups.

e) *Domain Specificity:* Emotion recognition can be context-dependent and domain-specific. Acknowledging this fact, we curated data from a wide range of domains, including social media, movie scripts, clinical records, and audiovisual recordings. This contrasting data collection approach ensures the robustness and adaptability of our model to different application scenarios. The inclusion of domain-specific data equips our model with versatility, enabling it to excel in specific contexts, such as mental health support, entertainment, or social interaction platforms.

Following the meticulous adherence to the stated metrics, data for all three modules (text, image, speech) was successfully gathered.

2) *Data Preprocessing and Analysis*

Data preprocessing played a critical role in the research, as it enabled us to harness the full potential of our distinct data set. This section delves into the steps we undertook to prepare the textual, image, and audio data for model training and analysis. By employing systematic data preprocessing techniques, we aimed to enhance the quality, balance, and representativeness of our dataset, thereby fortifying the foundation for building robust and accurate Emotion Recognition models.

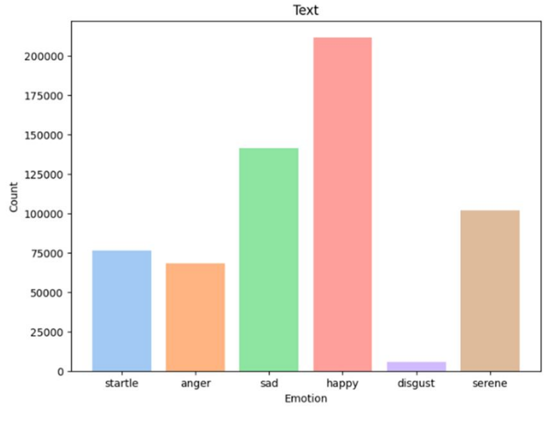
a) *Text*

The Text module constitutes a daunting dataset, comprising 605,033 records, with each record characterized by two essential columns: "text" and "emotion". The "text" column holds the textual data upon which our model undergoes rigorous training, while the "emotion" column designates the corresponding emotion category associated with each textual entry.

Acquiring a substantial corpus of text data presented its challenges, but our relentless efforts led us to discover and curate a plethora of textual sources that adhered to our established metrics and fulfilled our research objectives. While the process of data collection was not devoid of hurdles, we ensured the use of reliable and authentic sources to guarantee the integrity and relevance of our dataset.

Despite our dedicated endeavors, achieving a perfect data balance for all emotion classes posed a formidable challenge due to the natural variability and distribution of emotional expressions in real-world data sources. Nevertheless, we meticulously crafted the dataset with a focus on inclusion and variety, enabling it to surround a wide range of emotional patterns.

The distribution of the emotion classes in the Text module is as follows:



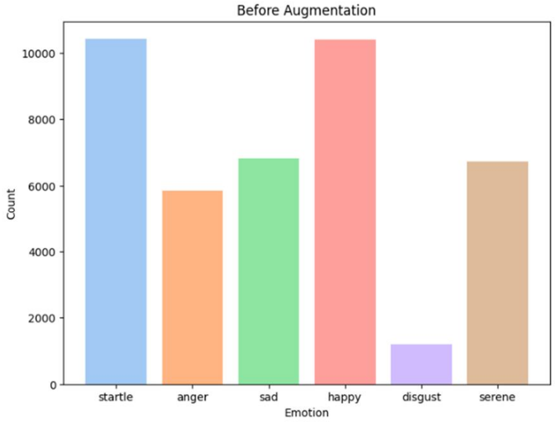
The dataset's classes have been encoded to represent the six distinctive emotion categories, with each category exhibiting a myriad distribution of records. In our subsequent analysis and model development, we scrupulously handled any class imbalance to ensure the unbiased performance of the Text Emotion Recognition model. The rich and inclusive text data gathered from legitimate sources sets the foundation for a robust and extensive investigation into emotion recognition through textual expressions.

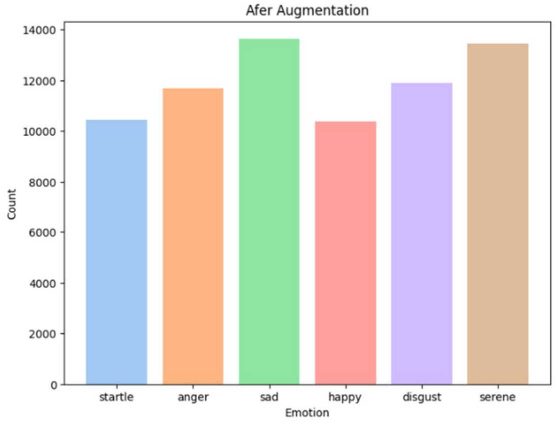
b) Image

For the Image module, we procured a valuable dataset provided by SUDARSHAN VAIDYA and JONATHAN OHEIX on Kaggle, which forms the backbone of our face-based emotion recognition study. These datasets comprise a total of about 52,000 images distributed among selected 6 classes, each meticulously labeled with their corresponding emotion categories.

To address any potential data imbalances and enhance the range of the image data, we leveraged data augmentation techniques. Augmentation allows us to generate synthetic variations of the existing images, thereby enriching the dataset and promoting better generalization of our model.

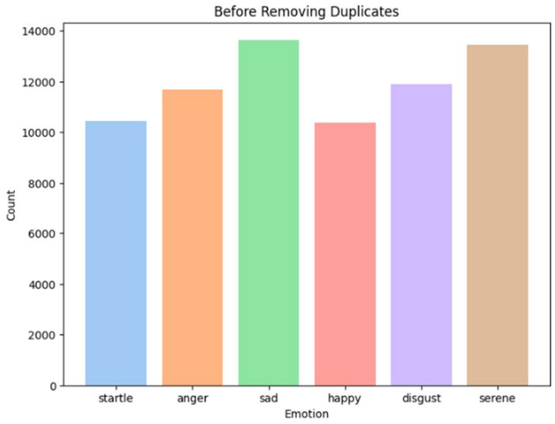
Before and after augmentation, the distribution of emotion classes in the Image module is as follows:

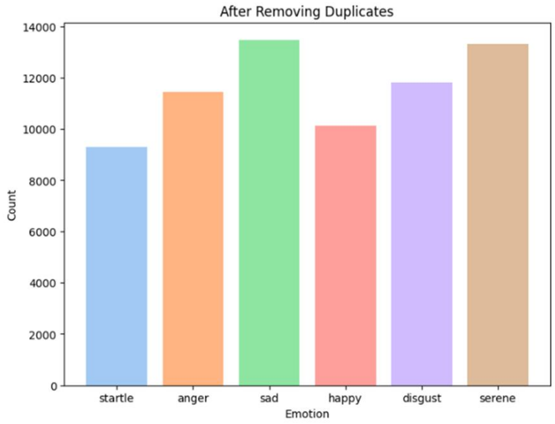




By employing data augmentation, we ensure that our Image Emotion Recognition model gains exposure to an assorted range of visual stimuli, thereby enhancing its capacity to discern and classify emotions accurately from images. The carefully curated and augmented image dataset paves the way for an in-depth exploration of emotion recognition through visual cues, setting the stage for our subsequent model development and analysis.

In addressing potential issues of duplicate images within our dataset, we implemented a rigorous approach leveraging the MD5 hashing algorithm from the “hashlib” library. This methodology enabled us to compute unique cryptographic hash values for each image, subsequently identifying and excluding duplicate instances before and after augmentation. This thorough curation process not only ensured the dataset's authenticity and quality but also bolstered the reliability of our subsequent research in emotion recognition.



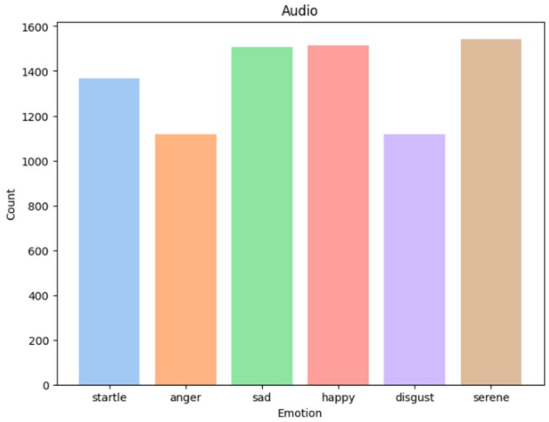


c) Audio

The audio module of the Emotion Recognition model integrates a range of audio files acquired from several reputable and well-established datasets. These datasets include the Ryerson Audio-Visual Database of Emotional Speech and Song RAVDESS, the Surrey Audio-Visual Expressed Emotion Database SAVEE, the Toronto Emotional Speech Set TESS, and the Crowd-sourced Emotional Multimodal Actors Dataset CREMA-D. By incorporating data from these multiple sources, we aimed to ensure a broad representation of emotional expressions and improve the model's generalization across different voice qualities, linguistic variations, and emotional nuances.

To ensure consistency and compatibility within our model, we have encoded these sundry emotion labels to our own six primary emotions, which are happy, sad, startle (encompassing emotions like fear and surprise), serene (including emotions like neutral and calm), angry, and disgust.

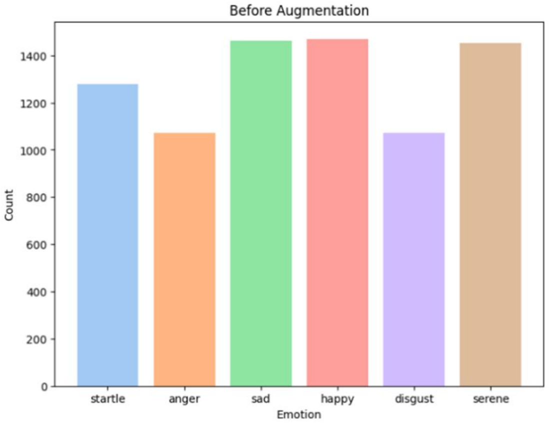
After data aggregation and encoding, the final distribution of emotional expressions in the audio dataset is as follows:

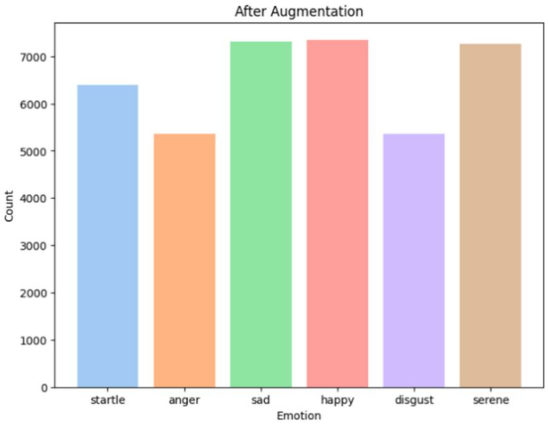


Combining the audio files from all discussed datasets resulted in an extensive and well-diversified dataset, ensuring a rich and representative sample of various emotional expressions across different speakers, emotions, and recording environments. The utilization of multiple datasets enabled us to create a more robust and adaptable Emotion Recognition model capable of recognizing emotions in different voices, accents, and acoustic conditions. Moreover, this integration of data from multiple sources enhanced the generalization ability of the model, enabling it to perform effectively on unseen or novel voice recordings. Our goal here was to leverage data acquisition and curation and create an audio module that can accurately recognize emotions in various real-world scenarios, making it suitable for a wide array of practical applications.

In the pursuit of completely enhancing our dataset for audio emotion recognition, we have diligently applied augmentation methodologies to our audio files. This augmentation process, designed to account for the variances in accents and pitch levels across the dataset, is founded upon four pivotal techniques: noise injection, temporal shifting, pitch modulation, and tempo alteration.

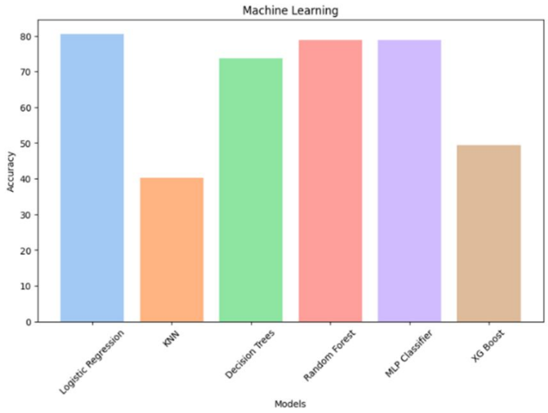
It is imperative to conduct a careful examination of the emotion class distribution within the Audio module below. At first glance, the distribution may appear seemingly unchanged before and after augmentation. However, upon closer scrutiny, it becomes apparent that there exists a nuanced distinction. This distinction arises from the fact that, during augmentation, four audio files are generated for each original file.





C. Implementation

The heart of our research lay in the development of advanced Emotion Recognition models that could effectively discern and classify emotions across different modalities. This section unraveled the intricacies of our model development process, from the selection of appropriate architectures and algorithms to the fine-tuning of hyperparameters. Leveraging cutting-edge technologies and best practices, we sought to create an ensemble of models that seamlessly processed textual, image, and audio inputs and produced insightful emotion predictions. Our modeling efforts were guided by the overarching objective of achieving state-of-the-art performance, thereby contributing to the advancement of Emotion Recognition technology and its transformative applications in various domains.



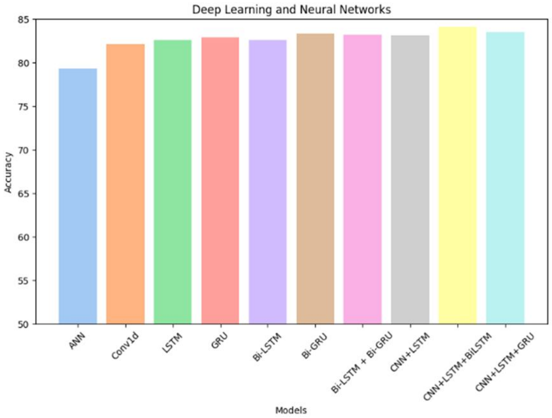
1) Text

We embarked on a thorough exploration of classical machine learning models and deep learning architectures for Text Emotion Recognition (TER). Our journey commenced with Logistic Regression, a simple yet interpretable model, extended for multiclass problems, serving as a foundational benchmark for TER. Moving forward, we delved into K-Nearest Neighbors (KNN), a non-parametric algorithm leveraging proximity in feature space, with varying 'k' values to optimize performance, yet it was not so efficient. Support Vector Machines (SVMs) followed, known for their capacity to capture intricate decision boundaries, where we experimented with different kernel functions and regularization parameters. Our exploration culminated in Decision Trees and Random Forest, harnessing ensemble learning for TER. We transitioned to the domain of deep learning with the Multi-Layer Perceptron (MLP) Classifier, showcasing the versatility of artificial neural networks in emotion recognition.

It becomes evident that the conventional machine learning models, hailing from the scikit-learn library, have exhibited rather lackluster performance. Notably, all these models, except for the Logistic Regression model, have struggled to provide satisfactory results. The Logistic Regression model, characterized by its simplicity and interpretability, has emerged as the surprising victor in this arena, boasting an accuracy score of 80.5% as visualized.

Shifting gears to the realm of deep learning, we introduced a series of models beginning with the classic Artificial Neural Network (ANN), which excels in capturing complex data relationships. We then ventured into the domain of 1D Convolutional Neural Networks (CNNs), well-suited for capturing local patterns within textual data.

Long Short-Term Memory (LSTM) networks were employed to capture sequential dependencies, while we expected Gated Recurrent Units (GRUs) to offer computational efficiency in comparison to LSTMs, it was not as such. The exploration of bidirectional models, including Bidirectional LSTMs (Bi-LSTMs) and Bidirectional GRUs (Bi-GRUs), highlighted the significance of contextual bidirectional processing. The performance of all the Neural Networks was almost like each other as per our research which persuaded us to try hybrid architectures combining these bidirectional models were designed to leverage their complementary strengths. Furthermore, the fusion of CNNs and LSTMs aimed to capture both local and global textual features, culminating in a complex model combining CNNs, LSTMs, and Bidirectional LSTMs. While the thought was that, it did not happen so. All the hybrid architectures well performed like the pure models.

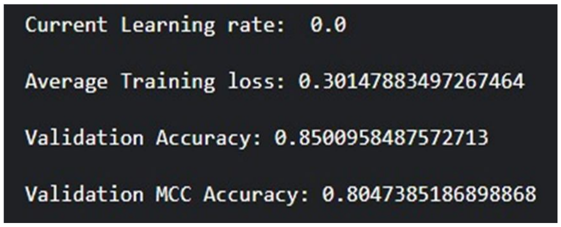


However custom-trained neural network models, each scrupulously designed and fine-tuned for the task of text emotion recognition, have demonstrated a level of performance that is consistent across the board. With an average accuracy rate of about 79%, these models have illustrated their competence in capturing intricate patterns within textual data and recognizing emotional nuances.

Model evaluation was conducted carefully, employing standard metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Visual representations of these results were presented, offering an intuitive comparative analysis of the models' performance in the intricate task of recognizing emotions from textual data. This exploration sets the stage for insightful discussions and conclusions regarding the optimal approaches for Text Emotion Recognition.

In the extensive evaluation of our text emotion recognition models, amidst these outcomes, the BERT base uncased model stands as a dominant contender that has eclipsed all other models in terms of accuracy. With an impressive accuracy score of 85%, the BERT model has secured its place as the final and most potent model in our ensemble. Its exceptional performance in the realm of text emotion recognition is a testament to the power of pre-trained contextual embeddings and transformer-based architectures.

The conclusive results of the BERT model upon completing its training epoch are as follows:



```
Current Learning rate:  0.0

Average Training loss: 0.30147883497267464

Validation Accuracy: 0.8500958487572713

Validation MCC Accuracy: 0.8047385186898868
```

The learning rate has converged to 0.0, signifying the culmination of the training process. The average training loss has been computed at 0.3014, indicating the degree of convergence achieved during the training phase. Notably, the validation accuracy, a critical metric for model evaluation, has reached an impressive 85.01%. Furthermore, the Matthews Correlation Coefficient (MCC) accuracy, another vital metric signifying the model's ability to handle imbalanced datasets, has demonstrated remarkable performance with an MCC accuracy score of 0.8047. These results collectively underscore the prowess of the BERT model in text emotion recognition and establish it as the pinnacle of our research endeavor.

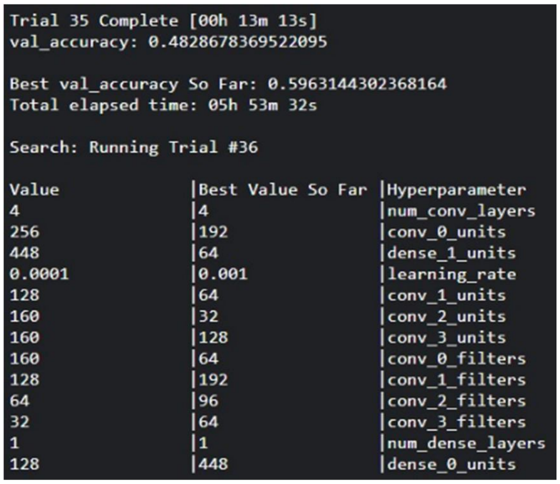
2) *Image*

In the initial phase of our research, we deployed a foundational Convolutional Neural Network (CNN) model aliased EmoNet by the team to gain insights into the intricate characteristics of our dataset. EmoNet underwent rigorous Hyperparameter Tuning, involving the exploration of approximately 100 varied combinations of hyperparameters within each layer and varying network depths. Despite these in-depth optimization efforts, EmoNet's peak accuracy reached a modest 59%. This outcome prompted us to broaden our approach, leading us to experiment with established standard models such as Resnet, VGG, InceptionV3, among others.

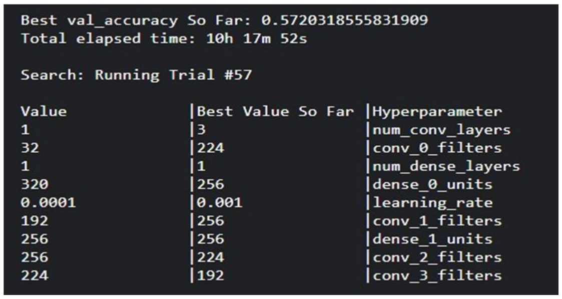
We carefully designed a data preprocessing and augmentation pipeline for the data discussed in section 3.2.2.2. This pipeline, a critical component of our image-based emotion recognition system, involves several key steps. Firstly, we established the batch size and image dimensions to efficiently manage and process our dataset. Subsequently, we employed the ImageDataGenerator from TensorFlow's Keras library, enabling us to perform real-time data augmentation and normalization during training. The dataset was divided into training and validation subsets with a 15% split to assess model performance effectively. During training, the data generator efficiently loads and preprocesses the images, ensuring they are appropriately rescaled and converted to grayscale. Furthermore, we ensured that the test dataset was processed consistently by implementing a separate test data generator that rescaled the test images.

Under the hyperparameter tuning, a systematic approach was adopted to optimize the architecture and configuration of EmoNet, our dedicated neural network model engineered for the precise recognition of emotions from facial expressions. To accomplish this intricate task, we harnessed the powerful capabilities of the Keras Tuner library, leveraging the RandomSearch algorithm to explore the hyperparameter space efficiently.

The results of hyperparameter tuning are as follows:



```
Trial 35 Complete [00h 13m 13s]
val_accuracy: 0.4828678369522095

Best val_accuracy So Far: 0.5963144302368164
Total elapsed time: 05h 53m 32s

Search: Running Trial #36

Value             |Best Value So Far |Hyperparameter
4                 |4                 |num_conv_layers
256               |192               |conv_0_units
448               |64                |dense_1_units
0.0001            |0.001             |learning_rate
128               |64                |conv_1_units
160               |32                |conv_2_units
160               |128               |conv_3_units
160               |64                |conv_0_filters
128               |192               |conv_1_filters
64                |96                |conv_2_filters
32                |64                |conv_3_filters
1                 |1                 |num_dense_layers
128               |448               |dense_0_units
```



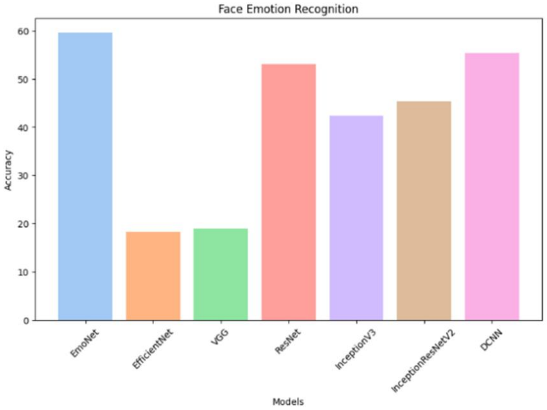
```
Best val_accuracy So Far: 0.5720318555831909
Total elapsed time: 10h 17m 52s

Search: Running Trial #57

Value             |Best Value So Far |Hyperparameter
1                 |3                 |num_conv_layers
32                |224               |conv_0_filters
1                 |1                 |num_dense_layers
320               |256               |dense_0_units
0.0001            |0.001             |learning_rate
192               |256               |conv_1_filters
256               |256               |dense_1_units
256               |224               |conv_2_filters
224               |192               |conv_3_filters
```

The optimization journey commenced with the definition of a crafted function called 'build_model.' This function served as the blueprint for assembling the neural network's architecture. Within this construct, an array of hyperparameters was precisely laid out for fine-tuning. To begin with, the number of convolutional layers was exposed to optimization, allowing EmoNet to adapt its depth dynamically based on the inherent complexity of the dataset. Each convolutional layer was characterized by its filter count, filter size, activation function, and pooling strategy. Additionally, critical components such as dropout and batch normalization layers were strategically inserted to bolster the model's ability to generalize and maintain training stability. Transitioning beyond the convolutional layers, a flattened layer facilitated the model's shift to densely connected layers, which were also amenable to hyperparameter tuning.

These dense layers enabled EmoNet to capture intricate patterns in the data. Moreover, the learning rate, a pivotal hyperparameter governing the model's learning process, underwent fine-tuning to regulate the rate of adaptation during training. Ultimately, the best hyperparameters emerged from this process, signifying the most optimal configuration for EmoNet.

While the Keras library offers a wide collection of pre-existing models, we undertook an approach to engineer several bespoke models from the ground up, aiming to extract the utmost potential from our dataset. Our model repertoire encompassed established architectures such as EfficientNet, Visual Geometry Group (VGG), ResNet, InceptionV3, InceptionResNetV2, and our custom-designed Deep Convolutional Neural Network (DCNN).

Upon subjecting these models to extensive training and evaluation, we observed noteworthy variations in their performance. EfficientNet and VGG models yielded initial accuracy rates of 19%, indicating a substantial room for improvement. In contrast, the ResNet architecture showcased a significant advancement, achieving a validation accuracy of 53%. Further, the InceptionV3 model exhibited a competitive accuracy of 42.4%, underscoring its potential in our context. The InceptionResNetV2 architecture delivered a commendable performance, attaining an accuracy rate of 45.3%. Finally, our custom designed DCNN model emerged as the frontrunner in terms of accuracy, achieving an impressive validation accuracy of 55.3%.



To conclude the Facial Emotion Recognition (FER), our findings highlight DCNN as the quintessential model we named it to be the EmoNet, exhibiting a laudable accuracy of approximately 55%. EmoNet, meticulously crafted through the orchestration of Conv2D, MaxPooling2D, Batch Normalization, and Dropout layers, followed by judiciously placed Dense layers, has emerged as the pinnacle of our explorations. In stark contrast, our investigations also unveil the less auspicious contenders, namely EfficientNet and VGG, both trailing behind with subpar accuracies, languishing at under 20%. This detailed assessment substantiates EmoNet's pre-eminence in the realm of FER models, while concurrently highlighting the limitations of alternative architectures in capturing nuanced facial emotions effectively.

3) Audio

In this section, we presented the implementation details of our SER system. The task is challenging due to the inherent variability in human speech patterns associated with different emotional states. Our work focuses on effectively capturing relevant features from audio data and employing a range of models to perform accurate emotion classification. These models were trained, validated, and tested to evaluate their performance in recognizing emotions from speech data discussed in Section 3.2.2.3.

To begin with, we conducted feature extraction from audio files using a combination of techniques including Mel-frequency cepstral coefficients (MFCCs), chroma features, and Mel-frequency spectrograms.

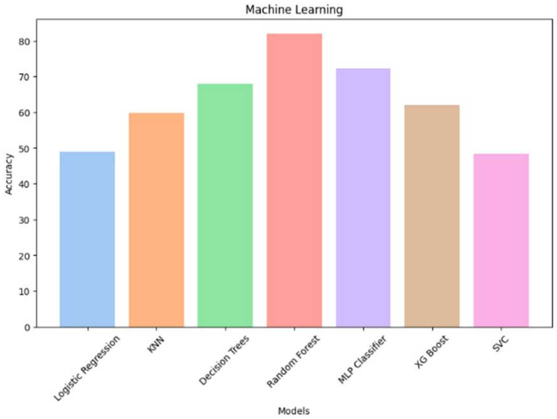
These features are known for their effectiveness in representing spectral characteristics and have been widely used in SER tasks.

1) *MFCCs:* We computed the MFCCs by applying the Librosa library to the audio data. A total of 40 MFCC coefficients by setting n_mfcc=40 was extracted from each audio file, capturing information related to the short-term power spectrum of the audio.
2) *Chroma Features:* Chroma features were extracted from the Short-Time Fourier Transform (STFT) of the audio. These features characterize the pitch content and harmony of the audio.
3) *Mel-frequency Spectrograms:* Mel-frequency spectrograms were derived to capture the spectral content of the audio signal in the Mel-frequency domain. This representation is particularly suitable for speech-related tasks.

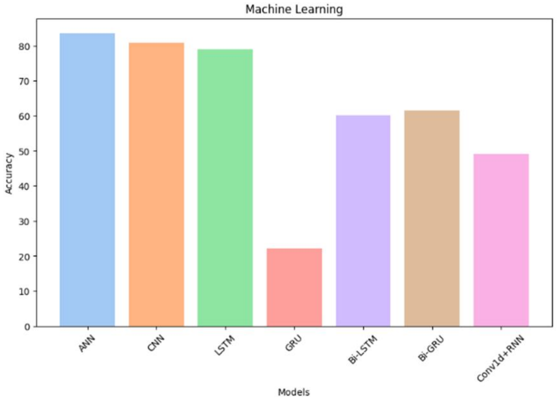
We obtained the emotional labels for each audio file by referencing a provided metadata file. Only a specific set of emotions, denoted as "observed_emotions," were selected for classification. We organized the dataset by loading audio files corresponding to each emotion category and extracting the features.

Our research entails an extensive exploration of machine learning and deep learning models for Speech Emotion Recognition (SER), with a focus on both traditional and neural network-based approaches just like the other modalities. The selected models are eclectic, covering a spectrum of complexities and architectural designs.

We initiated our investigation with Logistic Regression, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest. The model performance distribution coincidently looks like its normally distributed, however, to make it clear, it's just the order we used to visualize, and it does not imply anything.



Transitioning into deep learning, we explored the Dense Neural Network, characterized by densely connected layers, allowing for the learning of hierarchical representations. The Convolutional Neural Network (CNN) with 1D convolutional layers was employed to capture localized patterns within the audio data. Subsequently, recurrent neural networks (RNNs) entered the scene, with the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models addressing sequential data aspects. To capture bidirectional dependencies within the audio sequences, we leveraged the Bidirectional LSTM and Bidirectional GRU architectures. Here are the Performance Evaluation of DL models for SER:

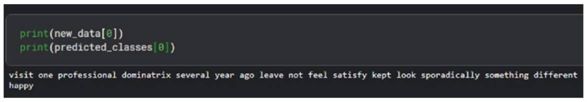


Our noteworthy observation was that as the complexity of the models increased, the classification accuracies exhibited a diminishing trend. This observation suggests that, for SER, the addition of model complexity does not always translate into performance gains, highlighting the importance of model selection and optimization in achieving the desired recognition accuracy.

Ultimately, in the realm of Speech Emotion Recognition (SER), our investigation was concluded with the identification of the most proficient machine learning model, which proved to be the Random Forest algorithm. Conversely, within the domain of deep learning, a remarkably effective yet elegantly simple solution emerged in the form of a fully connected neural network. These findings substantiate the notion that achieving optimal performance in SER involves a strategic amalgamation of various machine learning and deep learning techniques, where the adaptive and robust Random Forest stands out as an exemplar in the former category, while a streamlined yet potent fully connected network shines as a beacon of proficiency within the deep learning domain.

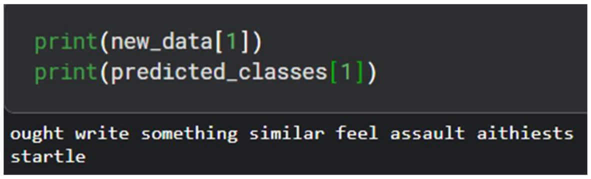
IV. RESULTS

A. Text Emotion Recognition



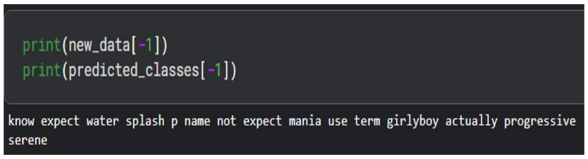
```
print(new_data[0])
print(predicted_classes[0])
```

```
visit one professional dominatrix several year ago leave not feel satisfy kept look sporadically something different
happy
```



```
print(new_data[1])
print(predicted_classes[1])
```

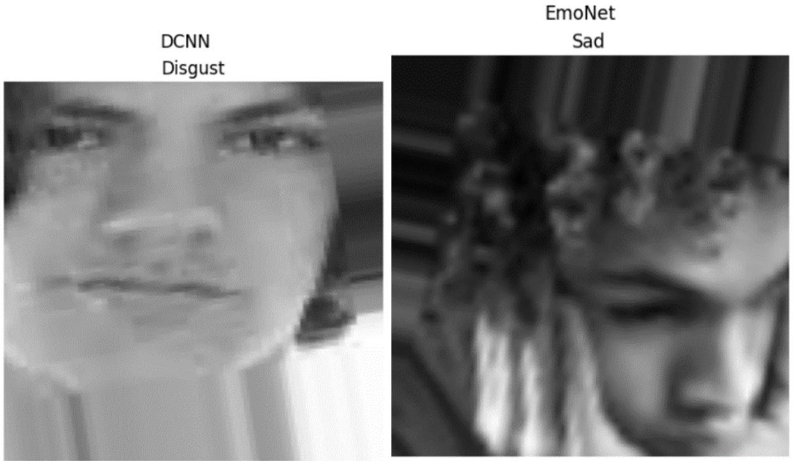
```
ought write something similar feel assault aithiests
startle
```



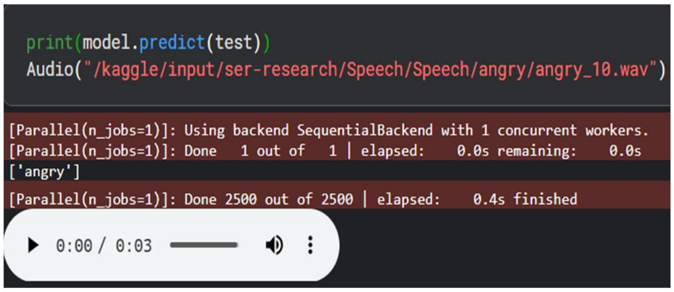
```
print(new_data[-1])
print(predicted_classes[-1])
```

```
know expect water splash p name not expect mania use term girlyboy actually progressive
serene
```

B. Face Emotion Recognition

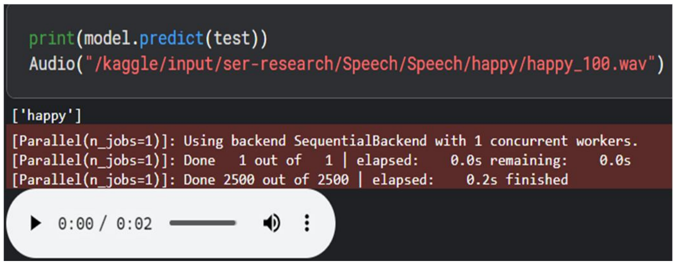


C. Speech Emotion Recognition



```
print(model.predict(test))
Audio("/kaggle/input/ser-research/Speech/Speech/angry/angry_10.wav")
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
['angry']
[Parallel(n_jobs=1)]: Done 2500 out of 2500 | elapsed:    0.4s finished
```



```
print(model.predict(test))
Audio("/kaggle/input/ser-research/Speech/Speech/happy/happy_100.wav")
```

```
['happy']
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    0.0s remaining:    0.0s
[Parallel(n_jobs=1)]: Done 2500 out of 2500 | elapsed:    0.2s finished
```

REFERENCES

[1] Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems, 115, 24-35. https://doi.org/10.1016/j.dss.2018.09.002

[2] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017), "End-to-End Multimodal Emotion Recognition using Deep Neural Networks",ArXiv, https://doi.org/10.1109/JSTSP.2017.2764438

[3] Tripathi, S., Tripathi, S., & Beigi, H. (2018), "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning", ArXiv, /abs/1804.05788

[4] W. Liu, J. -L. Qiu, W. -L. Zheng and B. -L. Lu, "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition," in IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 715-729, June 2022, doi: 10.1109/TCDS.2021.3071170.

[5] Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., & Li, X. (2019)," Learning Alignment for Multimodal Emotion Recognition from Speech", ArXiv, /abs/1909.05645.

[6] Salama, E. S., El-Khoribi, R. A., Shoman, M. E., & Wahby Shalaby, M. A. (2021), "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition", Egyptian Informatics Journal, 22(2), 167-176, https://doi.org/10.1016/j.eij.2020.07.005.

[7] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2019),"M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues", ArXiv,/abs/1911.05659.

[8] D. Priyasad, T. Fernando, S. Denman, S. Sridharan and C. Fookes, "Attention Driven Fusion for Multi-Modal Emotion Recognition," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 3227-3231, doi: 10.1109/ICASSP40776.2020.9054441.

[9] F. Lv, X. Chen, Y. Huang, L. Duan and G. Lin, "Progressive Modality Reinforcement for Human Multimodal Emotion Recognition from Unaligned Multimodal Sequences," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 2554-2562, doi: 10.1109/CVPR46437.2021.00258.

[10] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 2016, pp. 439-448, doi: 10.1109/ICDM.2016.0055.

[11] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[12] Hossain, M. S., & Muhammad, G. (2019)," Emotion recognition using deep learning approach from audio–visual emotional big data", Information Fusion, vol. 49, pp. 69-78, https://doi.org/10.1016/j.inffus.2018.09.008.

[13] Mellouk, W., & Handouzi, W. (2019), "Facial emotion recognition using deep learning: Review and insights", Procedia Computer Science, vol. 175, pp. 689-694, https://doi.org/10.1016/j.procs.2020.07.101.

[14] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2015), "Going Deeper in Facial Expression Recognition using Deep Neural Networks", ArXiv, https://doi.org/10.1109/WACV.2016.7477450.

[15] Han, Kun & Yu, Dong & Tashev, Ivan. (2014), "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine", Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 10.21437/Interspeech.2014-57.

[16] Y. Zhong, S. Qiu, X. Luo, Z. Meng and J. Liu, "Facial Expression Recognition Based on Optimized ResNet" 2020 2nd World Symposium on Artificial Intelligence (WSAI), Guangzhou, China, 2020, pp. 84-91, doi: 10.1109/WSAI49636.2020.9143287.

[17] https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset

[18] https://www.kaggle.com/datasets/sudarshanvaidya/random-images-for-face-emotion-recognition

[19] https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text/data

[20] https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text

[21] https://huggingface.co/datasets/dair-ai/emotion

AUTHORS PROFILE

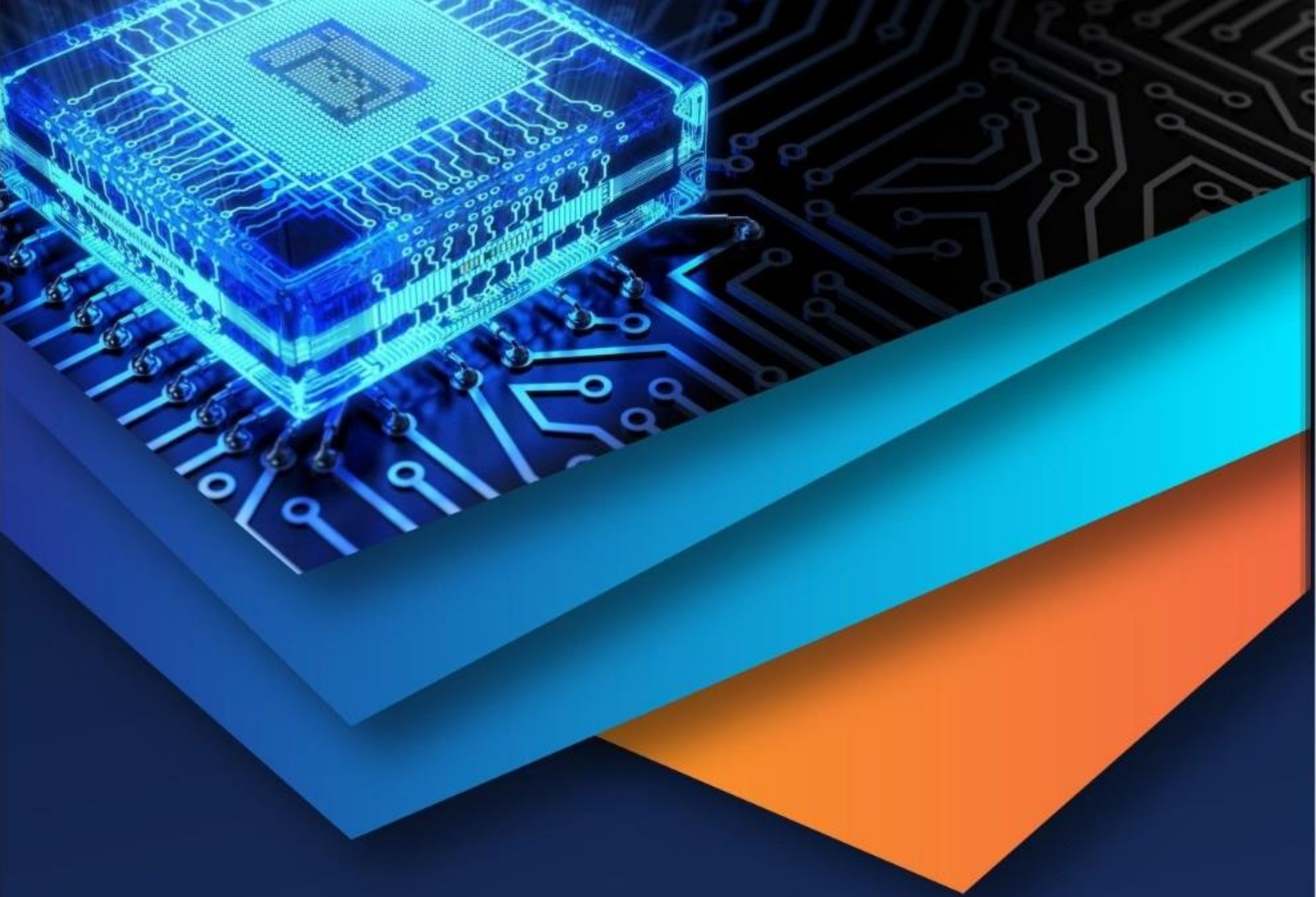
[1]: Dr. K. Shirisha Reddy, B. Tech (CSIT), M. Tech (SE), Ph. D (CSE). 18 Years of experience in teaching and currently working as a Head of the Department and an Associate Professor in Department of CSE [AI & ML] at Vignana Bharathi Institute of Technology.

[2]: Shreejit Cheela: Final Year Student in Department of CSE [AI & ML], VBIT.

[3]: Vignya Durvasula: Final Year Student in Department of CSE [AI & ML], VBIT.

OUR IMPLEMENTATION

https://bit.ly/ER_Resource



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)